ESSAYS ON BAYESIAN TIME SERIES AND VARIABLE SELECTION

A Dissertation

by

DEBKUMAR DE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Faming Liang
Co-Chair of Committee,	Bani K. Mallick
Committee Members,	Mohsen Pourahmadi
	Akhil Datta-Gupta
Department Head,	Simon Sheather

May 2014

Major Subject: Statistics

ABSTRACT

Estimating model parameters in dynamic model continues to be challenge. In my dissertation, we have introduced a Stochastic Approximation based parameter estimation approach under Ensemble Kalman Filter set-up. Asymptotic properties of the resultant estimates are discussed here. We have compared our proposed method to current methods via simulation studies. We have demonstrated predictive performance of our proposed method on a large spatio-temporal data.

In my other topic, we presented a method for simultaneous estimation of regression parameters and the covariance matrix, developed for a nonparametric Seemingly Unrelated Regression problem. This is a very flexible modeling technique that essentially performs a sparse high-dimensional *multiple predictor(p), multiple responses(q)* regression where the responses may be correlated. Such data appear abundantly in the fields of genomics, finance and econometrics. We illustrate and compare performances of our proposed techniques with previous analyses using both simulated and real multivariate data arising in econometrics and government.

To my parents and my wife, to whom I will remain forever indebted

ACKNOWLEDGEMENTS

First and foremost, I would like to take this opportunity to acknowledge my deepest gratitude to my advisors, Prof. Faming Liang and Prof. Bani K. Mallick. Without their continuous guidance, support and encouragement, this would not have been possible.

I would also like to thank Prof. Mohsen Pourahmadi for always patiently answering my silly questions, and giving me sound advice. Additionally, I would like to thank Prof. Akhil Datta-Gupta whose insightful comments have been extremely helpful.

Prof. Anindya Bhadra and Prof. Anirban Bhattacharya's inputs also have been of invaluable help and I thank them profusely.

Finally, I would like to take this formal opportunity to thank my parents and my wife for their continuous encouragement and unflinching support.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

In my dissertation, I worked on Bayesian time-series analysis and multivariate regression with graphical structure.

State-space models/Hidden Markov models have a long history in statistics literature. Introduced in 1960s, they have become more and more popular over time. In recent times, they have been widely used in diverse fields including, but not limited to, atmospheric science, petroleum engineering, finance, epidemiology. In his seminal work, R. E. Kalman (1960) introduced Kalman filter as an optimal solution for linear state-space model with Gaussian noise. Later Geir Evensen (1994) proposed Ensemble Kalman filter as an efficient solution for non-linear Gaussian state-space model.

Although most research in this area is focused on estimating the unobserved state variables, problems arising from estimating model parameters have also gained recognition among researchers. In the second section of my dissertation, we have proposed a Stochastic Approximation based parameter method for Gaussian state-space model. Stochastic approximation, introduced by Robbins and Monro (1951), has been widely used in sequentially approximating unique minimum(or maximum) to unimodal functions. The most beneficial quality of Stochastic approximation is that it enables quick "online" updating as and when new data becomes available. We have showed that the resultant estimates are asymptotically efficient. We have compared the performance of our proposed algorithm to that of another prevailing technique. section two also includes an application of our method to spatio-temporal atmospheric data.

In the third section of my dissertation, we introduce joint estimation technique for regression parameters as well as the variance-covariance parameters in case of non-linear Seemingly Unrelated Regression problem with a small sample size, relative to the number of model parameters to be estimated. Arnold Zellner (1962) introduced Seemingly Unre-

lated Regression(SUR) as an optimal solution to seemingly unrelated regression problem with correlated errors. Later non-parametric SUR methods are developed to solve the non-linear SUR problem. But all these models treat the variance-covariance parameter, $\mathbf{\Sigma}$, as nuisance parameter. In our proposed method, we explicitly model the covariance matrix by attaching a sparse graphical structure to our dependent variables. Gaussian graphical models are useful device to model high dimensional sparse graphical structure, regularly found in financial portfolio management, marketing, bioinformatics etc., by imposing conditional independence. This in turn provides efficiency and scalability in case of small sample, high dimensional problems. This section includes comparative studies using both simulated data and real world examples. Section four concludes my dissertation by discussing current limitations and future possibilities of my research.

2. AN ADAPTIVE ENSEMBLE KALMAN FILTER WITH PARAMETER ESTIMATION

2.1 Introduction

A mathematical model, through a system of equations involving multiple variables, tries to represent some physical phenomenon. Unfortunately, a single solution of the system of equations fails to capture the inherent uncertainty present in the physical world. Rather than considering a system of deterministic equations, a better approach would be to introduce uncertainty through probability density function for the involved model states. This helps us to asses multiple likely solutions together with the likelihood of realization of that solution. Based on the observed data, evaluation of the density function of the model solution is what is known as Data Assimilation or Inverse Problem.

The literature on data assimilation techniques is quite extensive. Some techniques try to map out an accurate path of the density function over time, where as others try to approximate the density function using moment estimates. In case of high dimensional models, representing the complete density path can become highly complicated, as well as computationally expensive. In such situations density functions are better represented through moment estimates and ensemble of model states. Among these various methods, some work well for simple linear process, but fails in case of nonlinear dynamics. Other methods are more suitable for non-linear problems, but are computationally expensive, and hence not optimal for linear dynamics.

Kalman filter was introduced by R. E. Kalman (1960) as a optimal solution for state estimation of dynamic linear systems under Gaussian noise. It has been extensively used in the fields of signal processing and navigation. Kalman Filter uses the second order moments, "integrated forward in time to predict error statistics" [i], which in turn are used

[i] Evensen (2009b)

to calculate minimum variance state estimates. "The filter is very powerful in several aspects: it supports estimations of past, present, and even future states, and it can do so even when the precise nature of the modeled system is unknown."[ii]. Despite of it's efficiency in solving discrete linear filtering problem, Kalman Filter suffers in case of high dimensional or non-linear models.

In case of high dimensional state vector in dynamical model, Kalman Filter(KF) suffers from computational issues. For a model with p unobserved state vector, the error covariance matrix consists of $p(p+1)/2$ unknowns at each time point. This restricts the use of Kalman Filter to moderately low dimensional problems. For a non-linear dynamic system, Kalman Filter solutions become suboptimal since in this case, the probability distribution of the unobserved states are not completely characterized by the second order moments. In this case the probability density of the model states follow what is known as Fokker-Planck equation (Kolmogorov's equation)(see Jazwinski (1970)). As a solution, Gelb (1974) developed Extended Kalman Filter(EKF), which follows the formulation of the Kalman filter with the Jacobian of dynamic matrix in place of the linear dynamic matrix. Although Extended Kalman filter was effective in many real life cases, but it is not an optimal estimator as it fails to account for the full non-linear dynamics. Extended Kalman Filter requires linearization for deriving error covariance evolution. But linearization leads to sub-optimal and unstable error covariance evolution. These issues can be resolved by involving higher order moments. But that would make it more computationally expensive.

Geir Evensen (1994) introduced Ensemble Kalman Filter(EnKF) as an approach to solve nonlinear state estimation. It was introduced as a way to solve the computational problem that arises in cases of KF and EKF. Ensemble Kalman filter is very popular for high-dimensional weather forecasting systems where models are extremely nonlinear in

[ii]Welch and Bishop (2007)

nature, initial states are very noisy and a large number of observations are available. The inherent ease of computational implementation and intuitive nature of it's formulation lead to it's widespread use. A detailed overview of the Ensemble Kalman Filter can be found in Evensen (2009b).

All the dynamic models discussed so far, namely, Kalman Filter, Extended Kalman Filter and Ensemble Kalman Filter, recursively updates model states based on fixed, known model parameters. Parameter estimation in dynamic models is different from usual methods. Here the focus is to formulate the problem in terms of joint density function of the model parameters and the unknown model states, conditioned on observed variables.

Parameter estimation in Ensemble Kalman filter setting has a long history, specially in the context of atmospheric science. Dee (1995) and Dee and da Silva (1999) provided on-line estimation of different parameters based on maximum likelihood approach. Mitchell and Houtekamer (2000) proposed an extension of this maximum likelihood based method. Ensemble Adjustment Kalman Filter was introduced by Anderson (2001) where he included unknown parameters in the state vector. Evensen (2009a) introduced Gaussian prior for the model parameters. Both Anderson (2001) and Evensen (2009a) are not applicable for variance parameters. Stroud and Bengtsson (2007) represented the beliefs about unknown parameters and the states by joint probability distribution. Their method is applicable in case of variance co-variance parameters. Unfortunately variance estimates turned out be unstable.

In my research, we propose a stochastic approximation based technique for parameter estimation. Stochastic approximation was introduced by Robbins and Monro (1951) and Kiefer and Wolfowitz (1952). It has since been developed into an important tool often used as a discrete time iterative stochastic optimization algorithm for solving online root finding problems(for detailed overview, see Lai (2003)). Under MCMC set-

up, Stochastic approximation has been extensively used to solve maximum likelihood estimation problems(see Younes (1988, 1999); Moyeed and Baddeley (1991); Gelfand and Banerjee (1998); Gu and Kong (1998); Delyon et al. (1999); Gu and Zhu (2001)) and monte carlo simulations(Liang et al. (2007)).

This chapter is organized as follows. In section 2, I have reviewed Kalman Filter, Extended Kalman Filter, Ensemble Kalman Filter and some existing parameter estimation techniques. In section 3, I give a basic review of Stochastic Approximation theory. Stochastic approximation in Ensemble Kalman Filter set-up is introduced in Section 4. In section 5, I have compared our algorithm with other competing through extensive simulation study based on linear spatio-temporal models as well as non-linear models. In section 6, I have introduced a modified parameter estimation method in case of limited data. Proposed method is applied on real data in section 7. Finally, in section 8, I conclude this chapter with discussion.

2.2 Review of various Kalman Filters and Parameter Estimation Techniques

Kalman filter is named after Rudolf E. Kalman, even though Thorvald Nicolai Thiele-Lauritzen (1981),Lauritzen (2002) and Peter SwerlingSwerling (1958) developed a similar algorithm earlier. Richard S. Bucy and Stanley F. Schmidt said to have contributed in developing the theoretical basis and the implementation of Kalman filter. Kalman Filter was used to solve the problem of trajectory estimation for the Apollo program, leading to its incorporation in the Apollo navigation computer. Early descriptions of Kalman Filter can be found in Kalman (1960) and Kalman and Bucy (1961).

Kalman filters have since been used in the implementation of U.S. Navy nuclear ballistic missile submarines navigation systems as well as in the guidance and navigation systems of various US missiles system such as Tomahawk missile and the Cruise Missile. It has also been heavily used in various space shuttle navigation system. For a more elaborate introductory discussion on Kalman Filter consult Sorenson (1970).

Let $\mathbf{y}_t$ and $\mathbf{x}_t$ be the p and q dimensional observation and hidden state variable at the time t. Let θ be set of unknown parameter. Figure (2.1) depicts the dynamic relation between $\mathbf{y}_t$ and $\mathbf{x}_t$. The generalized version of state-space model can be written as

$$\begin{aligned} \text{Measurement Model} \quad \mathbf{y}_t &\sim p(\mathbf{y}_t|\mathbf{x}_t, \theta) \\ \text{Dynamic Model} \quad \mathbf{x}_t &\sim p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta) \end{aligned}$$

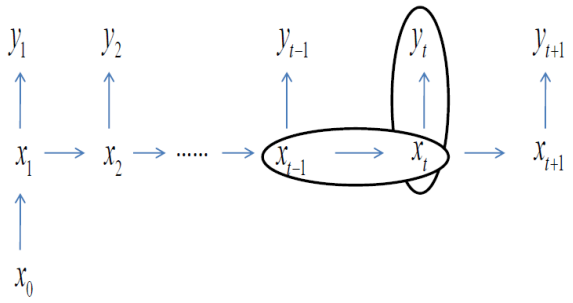

Figure 2.1: Dynamic Model. Given $\mathbf{x}_t$, $\mathbf{y}_t$ is independent of $\mathbf{y}_{1:t-1}$, $\mathbf{x}_{1:t-1}$.

2.2.1 Kalman Filter

Under the above set-up Kalman filter is an optimal variance-minimizing scheme which updates the state estimates with each new measurement when the measurement model and the dynamic model is assumed to be Gaussian with linear mean. Hence we have

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{u}_{t-1} + \mathbf{w}_{t-1}, \text{ where } \mathbf{w}_{t-1} \sim \mathrm{N}(0, \mathbf{Q}) \\ \mathbf{x}_0 &= \mathbf{A}_0 + \mathbf{w}_0, \\ \mathbf{y}_t &= \mathbf{H}\mathbf{x}_t + \mathbf{v}_t, \text{ where } \mathbf{v}_t \sim \mathrm{N}(0, \mathbf{R}), \end{aligned} \tag{2.1}$$

where $\mathbf{A}$ and $\mathbf{H}$ are the linear model operators, $\mathbf{A}_0$ is the initial condition with error $\mathbf{w}_0$, and $\mathbf{Q}$, $\mathbf{R}$ are known error variance matrices. In practise, both variance matrices as well

as the linear model operators can change over time. The random errors, $\mathbf{w}_t$ and $\mathbf{v}_t$ are assumed to be independent of each other. Here $\mathbf{u}_{t-1}$ is known optional control input.

2.2.1.1 Kalman Filter Algorithm

The discrete Kalman Filter state forecast step is the following

$$\mathbf{x}_t^f = \mathbf{A}\mathbf{x}_{t-1}^a + \mathbf{B}\mathbf{u}_{t-1}, \tag{2.2}$$

$$\mathbf{P}_t^f = \mathbf{A}\mathbf{P}_{t-1}^a\mathbf{A}' + \mathbf{Q}. \tag{2.3}$$

Here $\mathbf{x}_t^f$ is called the *a priori* state estimate or the state forecast at time t and $\mathbf{x}_t^a$ is the *a posteriori* or the updated state estimate at time step t. Similarly $\mathbf{P}_t^f$ and $\mathbf{P}_t^a$ are the a priori and a posteriori error covariance estimates. In Kalman Filter, our main goal is to find an way to get the a posteriori state estimate $\mathbf{x}_t^a$ as a linear combination of the a priori state estimate, $\mathbf{x}_t^f$ and the observed data, $\mathbf{y}_t$ at time t. The a posteriori state estimate $\mathbf{x}_t^a$ is given by

$$\mathbf{x}_t^a = \mathbf{x}_t^f + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}\mathbf{x}_t^f), \tag{2.4}$$

where $\mathbf{K}_t$ is called Kalman gain matrix. The Kalman gain matrix as well as the a posteriori error covariance matrix is derived as

$$\mathbf{K}_t = \mathbf{P}_t^f\mathbf{H}'(\mathbf{H}\mathbf{P}_t^f\mathbf{H}' + \mathbf{R})^{-1}, \tag{2.5}$$

$$\mathbf{P}_t^a = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\mathbf{P}_t^f. \tag{2.6}$$

After the updated state is estimated, the same process is repeated, to dynamically derive the a posteriori state estimate for step $(t+1)$. This dynamic nature of Kalman Filter makes it easy to implement.

In practical implementation of Kalman Filter, both the measurement model covariance $\mathbf{R}$ and the dynamic model covariance matrix $\mathbf{Q}$ are measures before the operating

the filter. Measuring $\mathbf{R}$ is relatively easier than measuring $\mathbf{Q}$. By tuning the estimates of the error covariance matrices, superior performance of the filter can be achieved. Sometime measurement noise as well as the dynamic noise do not stay constant. In those cases, appropriate choices of $\mathbf{R}_t$ and $\mathbf{Q}_t$ are used to perform the filtering.

2.2.2 Extended Kalman Filter

Kalman Filter provides optimal solution when both the measurement model and the dynamic model are linear. Problem arises if one of them turn out to be non-linear in nature. Extended Kalman Filter provides state estimation, by linearizing the mean and the covariance estimates in Kalman Filter. Suppose the current state-space model is given by,

$$\begin{aligned} \mathbf{x}_t &= f(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}, \theta) + \mathbf{w}_{t-1} \\ \mathbf{y}_t &= h(\mathbf{x}_t, \theta) + \mathbf{v}_t, \end{aligned} \tag{2.7}$$

where $f(\cdot)$ and $h(\cdot)$ are non-linear in nature and $\mathbf{w}_t$, $\mathbf{v}_t$ are Gaussian with covariance matrices $\mathbf{Q}$ and $\mathbf{R}$, respectively. θ is the set of model parameters. Also suppose $\mathbf{x}_0$, $\mathbf{w}_t$ and $\mathbf{v}_t$ are independent. Due to the inherent non-linearity of the system of equations, the distribution of the different variables no longer Gaussian in nature. EKF tries to approximate the optimality of KF through linearization. Here the state forecast step is given by

$$\mathbf{x}_t^f = f(\mathbf{x}_{t-1}^a, \mathbf{u}_{t-1}, \theta), \tag{2.8}$$

$$\mathbf{P}_t^f = \mathbf{A}_t \mathbf{P}_t^a \mathbf{A}_t^{'} + \mathbf{Q}, \tag{2.9}$$

and the state updates step is given by

$$\begin{aligned} \mathbf{x}_t^a &= \mathbf{x}_t^f + \mathbf{K}_t(\mathbf{y}_t - h(\mathbf{x}_t^f)), && (2.10) \\ \mathbf{K}_t &= \mathbf{P}_t^f \mathbf{C}_t^{'}(\mathbf{C}_t \mathbf{P}_t^f \mathbf{C}_t^{'} + \mathbf{R})^{-1}, && (2.11) \\ \mathbf{P}_t^a &= (\mathbf{I} - \mathbf{K}_t)\mathbf{C}_t \mathbf{P}_t^f, \end{aligned}$$

where the Jacobians $\mathbf{A}_t$ and $\mathbf{C}_t$ are defined as $\mathbf{A}_t = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{x}_{t-1}^a}$ and $\mathbf{C}_t = \frac{\partial h(\mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{x}_{t-1}^a}$. See Jazwinski (1970), Gelb (1974) for more details.

Like the Kalman Filter, Extended Kalman Filter is also recursive in nature. But unlike KF, EKF has a tendency to diverge outside few restrictive cases. This makes the model states unestimable. Julier et al. (1995) proposed some variation of EKF that maintains Normal distribution, even under non-linear transformation.

2.2.3 Ensemble Kalman Filter

Geir Evensen (1994) proposed Ensemble Kalman Filter as a solution to problems arising from the non-linear state-space models. EnKF provides a Monte Carlo solution to the problems arising from high dimensional non-linear state-space models. Computational affordability of the EnKF has lead to it's popularity. Ensemble Kalman filter applies Monte Carlo method through an ensemble of states to solve the Fokker-Planck equation. These ensemble of model states are used to predict the error statistics. And the covariance matrix is replaced by the sample covariance computed from the ensemble.

Let us start by defining an ensemble of size m forecasted states at time step t,

$$\mathbf{X}_t^f = [\mathbf{x}_t^{f_1} \dots \mathbf{x}_t^{f_m}] \in \mathbb{R}^{q*m},$$

where $\mathbf{x}_t^{f_i}$ is the i$^{\text{th}}$ forecasted state ensemble member at time t. Similarly, we denote the

predicted ensemble observation at time t as

$$\mathbf{Y}_t^f = [\mathbf{y}_t^{f_1} \dots \mathbf{y}_t^{f_m}] \in \mathbb{R}^{p*m}.$$

The ensemble means are defined as

$$\bar{\mathbf{x}}_t^f = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_t^{f_i} \in \mathbb{R}^q \quad \text{and} \quad \bar{\mathbf{y}}_t^f = \frac{1}{m}\sum_{i=1}^{m} \mathbf{y}_t^{f_i} \in \mathbb{R}^p.$$

We can then define the state error matrix at time t as,

$$\mathbf{E}_{\mathbf{x}_t}^f = [\mathbf{x}_t^{f_1} - \bar{\mathbf{x}}_t^f \dots \mathbf{x}_t^{f_m} - \bar{\mathbf{x}}_t^f] \in \mathbb{R}^{q*m},$$

and the corresponding measurement error matrix as

$$\mathbf{E}_{\mathbf{y}_t}^f = [\mathbf{y}_t^{f_1} - \bar{\mathbf{y}}_t^f \dots \mathbf{y}_t^{f_m} - \bar{\mathbf{y}}_t^f] \in \mathbb{R}^{p*m}.$$

Unlike the Kalman Filter, where we can calculate the exact error covariance matrices, in EnKF we estimate the true covariance matrices by the sample covariance matrices, defined as

$$\begin{aligned}
\hat{\mathbf{P}}_{xx_t}^f &= \frac{1}{m-1}\mathbf{E}_{\mathbf{x}_t}^f \mathbf{E}_{\mathbf{x}_t}^{f'}, \\
\hat{\mathbf{P}}_{yy_t}^f &= \frac{1}{m-1}\mathbf{E}_{\mathbf{y}_t}^f \mathbf{E}_{\mathbf{y}_t}^{f'}, \\
\hat{\mathbf{P}}_{yx_t}^f &= \frac{1}{m-1}\mathbf{E}_{\mathbf{y}_t}^f \mathbf{E}_{\mathbf{x}_t}^{f'}.
\end{aligned}$$

Next we obtain the analyzed or a posteriori ensemble state estimate $\mathbf{x}_t^{f_i}$ as

$$\mathbf{x}_t^{a_i} = \mathbf{x}_t^{f_i} + \hat{\mathbf{K}}_t(\mathbf{y}_t + \mathbf{v}_t^{f_i} - h(\mathbf{x}_t^{f_i}, \theta)), \quad \mathbf{v}_t^{f_i} \sim \mathrm{N}(0, \mathbf{R}), \tag{2.12}$$

where the sample Kalman gain matrix, $\hat{\mathbf{K}}_t$, is given by

$$\hat{\mathbf{K}}_t = \hat{\mathbf{P}}^f_{yx_t}(\hat{\mathbf{P}}^f_{yy_t})^{-1}.$$

The final step, forecast step, involves defining an ensemble of m forecasted states for time $(t+1)$ as,

$$\mathbf{x}^{f_i}_{t+1} = f(\mathbf{x}^{f_i}_t, \mathbf{u}_t, \theta) + \mathbf{w}^{f_i}_t, \quad \mathbf{w}^{f_i}_t \sim \mathrm{N}(0, \mathbf{Q}). \tag{2.13}$$

The sample error covariance matrix computed from $\mathbf{w}^{f_i}_t$ and $\mathbf{v}^{f_i}_t$ converges to $\mathbf{Q}$ and $\mathbf{R}$ respectively as $m \to \infty$. To sum up the steps so far, the Ensemble Kalman filter works as follows:

- **Forecast**

$$\begin{aligned}
\mathbf{x}^{f_i}_t &= f(\mathbf{x}^{a_i}_{t-1}, \mathbf{u}_{t-1}, \theta) + \mathbf{w}^i_{t-1}, \\
\mathbf{y}^{f_i}_t &= h(x^{f_i}_t, \theta) + \mathbf{v}^i_t, \\
\mathbf{E}^f_{\mathbf{x}_t} &= [\mathbf{x}^{f_1}_t - \bar{\mathbf{x}}^f_t \ldots \mathbf{x}^{f_m}_t - \bar{\mathbf{x}}^f_t], \\
\mathbf{E}^f_{\mathbf{y}_t} &= [\mathbf{y}^{f_1}_t - \bar{\mathbf{y}}^f_t \ldots \mathbf{y}^{f_m}_t - \bar{\mathbf{y}}^f_t], \\
\hat{\mathbf{P}}^f_{xx_t} &= \frac{1}{m-1}\mathbf{E}^f_{\mathbf{x}_t}\mathbf{E}^{f'}_{\mathbf{x}_t}, \quad \hat{\mathbf{P}}^f_{yy_t} = \frac{1}{m-1}\mathbf{E}^f_{\mathbf{y}_t}\mathbf{E}^{f'}_{\mathbf{y}_t}, \quad \hat{\mathbf{P}}^f_{yx_t} = \frac{1}{m-1}\mathbf{E}^f_{\mathbf{y}_t}\mathbf{E}^{f'}_{\mathbf{x}_t}.
\end{aligned}$$

- **State Update**

$$\begin{aligned}
\hat{\mathbf{K}}_t &= \hat{\mathbf{P}}^f_{\mathbf{x}\mathbf{y}_t}(\hat{\mathbf{P}}^f_{\mathbf{y}\mathbf{y}_t})^{-1}, \\
\mathbf{x}^{a_i}_t &= \mathbf{x}^{f_i}_t + \hat{\mathbf{K}}_t(\mathbf{y}_t + \mathbf{v}^i_t - h(\mathbf{x}^{f_i}_t, \theta)).
\end{aligned}$$

EnKF is much less computationally expensive than EKF, although the computational cost in the forecast step increases with m. Also, EnKF fails solve the update equation in case of non-Gaussian distribution functions. To summarize, Ensemble Kalman Filter pro-

vided a computationally efficient online state estimation technique in case of a non-linear Gaussian state space model. Another big limitation of EnKF is it introduces spurious correlation, through ensembles, which leads to shrinkage in the ensemble variance, which in turn leads to filter divergence. Using covariance inflation(see Anderson and Anderson (1999), Anderson (2007)) and localization(Hamill et al. (2001),Houtekamer and Mitchell (2001)) one can avoid filter divergence.

2.2.4 Parameter Estimation

Various Kalman Filter methods discussed so far are based on known model parameter. Problem arises when the model parameters are unknown. Main focus of Parameter estimation in dynamic model is to jointly estimate model parameters together with unobserved states. In Dee (1995) and Dee and da Silva (1999) introduced Maximum Likelihood based sequential and offline parameter estimation method. Later Mitchell and Houtekamer (2000) extended their maximum likelihood based technique to estimate variance parameters under EnKF set-up. Anderson (2001) estimated unknown model parameters by augmenting them with the state variables. Evensen (2009a) assumed Gaussian distribution for the model parameters and used augmentation method to estimate model parameters. In their paper, DelSole and Yang (2010) shown that the augmentation approach fails to give good estimates of variance covariance parameters(stochastic parameters). Stroud and Bengtsson (2007) introduced a fully Bayesian parameter estimation method. Although their method can estimate variance parameters under Gaussian error, their method becomes computationally expensive in case of model parameters for which no closed form posterior distribution exist.

2.3 Stochastic Approximation

Stochastic approximation(SA) was introduced by Robbins and Monro (1951). Consider the following problem of finding an unique root θ to the regression problem

$$y_n = M(x_n) + \epsilon_n, \qquad n = 1, 2, \ldots, \tag{2.14}$$

where ϵ_n is the unobserved random error. Here x_n is used as sequential approximation to θ. In case of a deterministic problem, i.e. $\epsilon_n \equiv 0$, Newton's method would imply $x_{n+1} = x_n - \frac{y_n}{M'(x_n)}$, under the assumption $M'(\theta) \neq 0$. Extending that to (2.14), we get

$$x_{n+1} = x_n - \frac{M(x_n) + \epsilon_n}{M'(x_n)} = x_n - \frac{M(x_n)}{M'(x_n)} - \frac{\epsilon_n}{M'(x_n)}.$$

Looking at the above equation, it is clear that for x_n to converge to θ, we need to have $\epsilon_n \to 0$, which in case of most of the random errors is not a valid assumption. To solve this problem, Robbins and Monro (1951) suggested to following sequential solution

$$x_{n+1} = x_n - \gamma_n y_n, \tag{2.15}$$

where $\{\gamma_n\}_n > 0$ satisfy $\sum_{n=0}^{\infty} \gamma_n = \infty$, $\sum_{n=0}^{\infty} \gamma_n^2 < \infty$, under the assumption that for $x_n > \theta$, $M(x_n) > 0$ and $x_n < \theta$, $M(x_n) < 0$. $\sum_{n=0}^{\infty} \gamma_n^2 < \infty$ ensures that $(x_n - \theta)$ converges almost surely and the condition $\sum_{n=0}^{\infty} \gamma_n = \infty$ ensures $x_n - \theta$ converges to zero.

Kiefer and Wolfowitz (1952) extended the Robbins and Monro (1951) algorithm to recursively provide the minimum (or maximum) of an unimodal function. They proposed

$$x_{n+1} = x_n - \gamma_n \Delta(x_n), \tag{2.16}$$

as a successive solution to the the problem $\frac{dM}{dx} = 0$. The ideas of stochastic approximation was later extended and developed in various areas of optimization and stochastic system. Based of stochastic approximation, many recursive algorithms were later developed. For a detailed discussion on the convergence and the asymptotic properties of stochastic approximation, as well as later developments in this area, see Lai (2003).

2.4 Estimating Parameter Using Stochastic Approximation in Ensemble Kalman Filter

In our research we have introduced a parameter estimation technique based on Stochastic Approximation. Consider the dynamic model given by

$$\begin{aligned} \mathbf{x}_t &= f(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}, \alpha) + \mathbf{w}_{t-1}, \qquad \mathbf{w}_{t-1} \sim \mathrm{N}(0, \mathbf{Q}(\eta_x)) \\ \mathbf{y}_t &= \mathbf{H}(\beta)\mathbf{x}_t + \mathbf{v}_t, \qquad\qquad\quad \mathbf{v}_t \sim \mathrm{N}(0, \mathbf{R}(\eta_y)), \end{aligned} \tag{2.17}$$

where $\theta = (\alpha, \beta, \eta_x, \eta_y)$ is the set of unknown parameters to be estimated. Here both η_x and η_y are time invariant parameter vectors. The nonlinear propagator $f(\cdot)$ contains the time-invariant model parameter vector α, and the linear propagator $\mathbf{H}(\cdot)$ contains the time-invariant model parameter vector β. The propagator $\mathbf{H}(\cdot)$ relates the state variables to the measured variables and yields the expected value of the prediction given the model states and parameters.

From (2.17), the conditional distribution of $\mathbf{y}_t$, given the state $\mathbf{x}_{t-1}$, follows a multivariate normal distribution $\mathrm{N}_p(\mathbf{H}(\beta)f(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}, \alpha), \mathbf{H}(\beta)\mathbf{Q}(\eta_x)\mathbf{H}'(\beta) + \mathbf{R}(\eta_y))$ with the density function

$$p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta) = (2\pi)^{-\frac{p}{2}}|\mathbf{\Sigma}(\theta)|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{y}_t - \boldsymbol{\mu}(\theta))'\mathbf{\Sigma}(\theta)^{-1}(\mathbf{y}_t - \boldsymbol{\mu}(\theta))],$$

where $\boldsymbol{\mu}(\theta) = \mathbf{H}(\beta) f(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}, \alpha)$ and $\boldsymbol{\Sigma}(\theta) = \mathbf{H}(\beta)\mathbf{Q}(\eta_x)\mathbf{H}'(\beta) + \mathbf{R}(\eta_y)$. Define,

$$\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \theta) = \frac{\partial \log p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta)}{\partial \theta} \text{ and } \Phi(\mathbf{y}_t, \mathbf{X}_{t-1}^a, \theta) = \frac{1}{m}\sum_{i=1}^{m} \phi(\mathbf{y}_t, \mathbf{x}_{t-1}^{a_i}, \theta),$$

where $\mathbf{X}_t^a = [\mathbf{x}_{t-1}^{a_1}, \cdots, \mathbf{x}_{t-1}^{a_m}]$.

Fix an arbitrary initial value θ_0, and denote by $\theta_t = (\alpha_t, \beta_t, \eta_{x,t}, \eta_{y,t})$ the estimate at iteration t which serves at the t^{th} approximation to θ. Let γ_t be a positive and non-increasing sequence satisfying the conditions

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \text{ and } \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

Then our method can be summarized as follows.

Algorithm 1(Adaptive EnKF with Parameter Estimation)

1. **Forecast Step** Set

$$\begin{aligned} \mathbf{x}_t^{f_i} &= f(\mathbf{x}_{t-1}^{a_i}, \mathbf{u}_{t-1}, \alpha_{t-1}) + \mathbf{w}_t^i, \\ \mathbf{y}_t^{f_i} &= \mathbf{H}(\beta_{t-1})\mathbf{x}_t^{f_i} + \mathbf{v}_t^i, \end{aligned}$$

   where $\mathbf{w}_t^i \sim \mathrm{N}(0, \mathbf{Q}(\eta_{x,t}))$ and $\mathbf{v}_t^i \sim \mathrm{N}(0, \mathbf{R}(\eta_{y,t}))$. Calculate $\hat{\mathbf{P}}_{xx_t}^f$.

2. **State Update**

$$\begin{aligned} \hat{\mathbf{K}}_t &= \hat{\mathbf{P}}_{xx_t}^f \mathbf{H}(\beta_{t-1})[\mathbf{H}(\beta_{t-1})\hat{\mathbf{P}}_{xx_t}^f \mathbf{H}'(\beta_{t-1}) + \mathbf{R}(\eta_{y,t})]^{-1} \\ \mathbf{x}_t^{a_i} &= \mathbf{x}_t^{f_i} + \hat{\mathbf{K}}_t(\mathbf{y}_t - \mathbf{y}_t^{f_i}). \end{aligned}$$

3. **Parameter Update**

$$\theta_t = \theta_{t-1} + \gamma_t \Phi(\mathbf{y}_t, \mathbf{X}^a_{t-1}, \theta_t),$$

We can also consider a projected version of this algorithm, replacing Parameter Update step by

$$\theta_t = \pi_{\bar{\Theta}}(\theta_{t-1} + \gamma_t \Phi(\mathbf{y}_t, \mathbf{X}^a_{t-1}, \theta)),$$

where $\bar{\Theta}$ is a bounded set of the form $\bar{\Theta} = \{x : q_i(x) \leq 0, i = 1, \ldots, s\}$, $q_i(\cdot)$'s are continuously differentiable, $\bar{\Theta}$ is the closure of its interior, and $\pi_{\bar{\Theta}}(z)$ denotes any closet point in $\bar{\Theta}$ to z. The projection step ensures that $\{\theta_t\}$ can be included in a bounded set.

2.4.1 Large Sample Asymptotic for the Ensemble Kalman filter

In this section, we study the large sample asymptotic for the EnKF algorithm with a fixed value of θ. In this case, the EnKF algorithm can be described as follows, where, for simplicity, the parameters are depressed.

Algorithm 2(EnKF with fixed parameters)

1. **Forecast Step** Set

$$\begin{aligned} \mathbf{x}^{f_i}_t &= f(\mathbf{x}^{a_i}_{t-1}, \mathbf{u}_{t-1}) + \mathbf{w}^i_t, \\ \mathbf{y}^{f_i}_t &= \mathbf{H}\mathbf{x}^{f_i}_t + \mathbf{v}^i_t, \end{aligned}$$

where $\mathbf{w}^i_t \sim \mathrm{N}(0, \mathbf{Q})$ and $\mathbf{v}^i_t \sim \mathrm{N}(0, \mathbf{R})$. Calculate $\hat{\mathbf{P}}^f_{xx_t}$.

2. **State Update**

$$\begin{aligned}\hat{\mathbf{K}}_t &= \hat{\mathbf{P}}^f_{xx_t}\mathbf{H}[\mathbf{H}\hat{\mathbf{P}}^f_{xx_t}\mathbf{H}' + \mathbf{R}]^{-1} \\ \mathbf{x}^{a_i}_t &= \mathbf{x}^{f_i}_t + \hat{\mathbf{K}}_t(\mathbf{y}_t - \mathbf{y}^{f_i}_t).\end{aligned}$$

To study the asymptotic behavior of the empirical probability distributions

$$\mu^{m,f}_k = \frac{1}{m}\sum_{i=1}^{m}\delta_{\mathbf{x}^{f_i}_k} \text{ and } \mu^{m,a}_k = \frac{1}{m}\sum_{i=1}^{m}\delta_{\mathbf{x}^{a_i}_k},$$

of the forecast elements and analysis elements, we follow Le Gland et al. (2009) to consider the decoupled EnKF algorithm:

Algorithm 3(Decoupled EnKF)

1. **Forecast Step** Set

$$\begin{aligned}\bar{\mathbf{x}}^{f_i}_t &= f(\bar{\mathbf{x}}^{a_i}_{t-1}, \mathbf{u}_{t-1}) + \mathbf{w}^i_t, \\ \bar{\mathbf{y}}^{f_i}_t &= \mathbf{H}\bar{\mathbf{x}}^{f_i}_t + \mathbf{v}^i_t,\end{aligned}$$

   where $\mathbf{w}^i_t \sim \mathrm{N}(0, \mathbf{Q})$ and $\mathbf{v}^i_t \sim \mathrm{N}(0, \mathbf{R})$.

2. **State Update**

$$\bar{\mathbf{x}}^{a_i}_t = \bar{\mathbf{x}}^{f_i}_t + \bar{\mathbf{K}}_t(\mathbf{y}_t - \bar{\mathbf{y}}^{f_i}_t),$$

   where

$$\bar{\mathbf{K}}_t = \bar{\mathbf{P}}^f_{xx_t}\mathbf{H}[\mathbf{H}\bar{\mathbf{P}}^f_{xx_t}\mathbf{H}' + \mathbf{R}]^{-1},$$

   with $\bar{\mathbf{P}}^f_{xx_t} = \mathbf{E}(\bar{\mathbf{x}}^{f_i}_t - \mathbf{E}(\bar{\mathbf{x}}^{f_i}_t))(\bar{\mathbf{x}}^{f_i}_t - \mathbf{E}(\bar{\mathbf{x}}^{f_i}_t))'$.

Note that for this decoupled algorithm, the resulting elements in the ensembles $\bar{\mathbf{x}}^a_t$ and $\bar{\mathbf{x}}^f_t$ are iid, respectively.

Let $\bar{\mu}_t^f(dx)$ and $\bar{\mu}_t^a(dx)$ be the probability distributions of $\bar{\mathbf{x}}_t^{f_i}$ and $\bar{\mathbf{x}}_t^{a_i}$ respectively. Let

$$\mu_t^-(dx) = P(\mathbf{x}_t \in dx|\mathbf{y}_{1:t-1}) \text{ and } \mu_t(dx) = P(\mathbf{x}_t \in dx|\mathbf{y}_{1:t}), \tag{2.18}$$

where $\mathbf{y}_{1:t} = (t_1, \ldots, y_m)$, be the Bayesian prediction distribution and the Bayesian filter distribution, respectively. Regarding the relationship between these distributions, Le Gland et al. (2009) established the following lemma.

Lemma 2.4.1. *(Le Gland et al. (2009); Lemma 2.1, p.6)*

(i) For linear systems with additive Gaussian white noises and with Gaussian initial condition, $\bar{\mu}_t^f$ and $\bar{\mu}_t^a$ coincide with the Gaussian distribution associated with the Bayesian predictor μ_t^- and the Bayesian filter μ_t, respectively.

(ii) For nonlinear systems with additive Gaussian white noises and with non-necessarily Gaussian initial condition, $\bar{\mu}_t^f$ and $\bar{\mu}_t^a$ differ from the Bayesian predictor μ_t^- and the Bayesian filter μ_t, respectively.

Definition 2.4.1. *(Lip(C,κ) function) A function r(x) is said to be a Lip(C,κ) function if there exists a constant $C > 0$ and a constant $\kappa > 0$ such that*

$$|r(x) - r(x')| \leq C|x - x'|(1 + |x|^\kappa + |x'|^\kappa), \tag{2.19}$$

for any $x, x' \in \mathbb{R}^{d_x}$; that is, r(x) is a locally Lipschitz continuous function with at most polynomial growth at infinity.

Lemma 2.4.2. *(Le Gland et al. (2009); Lemma 5.1, p.10) Consider a stationary dynamic system as described in (2.17), where the drift function $f(\cdot)$ is a Lip(L',κ') function for some constants $L' > 0$ and $\kappa > 0$. Let $\psi(x)$ be a Lip(L,κ) function for some constants L > 0 and kappa > 0. If the initial ensemble $\mathbf{X}_0^f$ has finite moments of any order, then for*

any value of θ *and any value of* $t > 0$,

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m}\psi(\mathbf{x}_t^{f_i}) &\rightarrow \int \psi(x)\bar{\mu}_t^f(dx) \ a.s., \\ \frac{1}{m}\sum_{i=1}^{m}\psi(\mathbf{x}_t^{a_i}) &\rightarrow \int \psi(x)\bar{\mu}_t^a(dx) \ a.s., \end{aligned}$$

as $m \rightarrow \infty$

2.4.2 Stochastic Approximation

We consider stochastic approximations of the form

$$\theta_t = \theta_{t-1} + \gamma_t B(\theta_{t-1}, \xi_t), \tag{2.20}$$

where the evolution of ξ_t can depend on θ_t in the sense that, in general,

$$P(\xi_t \in A | \xi_i, \ \ i \leq t-1) \neq P(\xi_t \in A | \theta_i, \xi_i, \ \ i \leq t-1).$$

We also treat the following projected version of (2.20). Let $\bar{\Theta} = \{x : q_i(x) \leq 0, i = 1, \ldots, s\}$, $q_i(\cdot)$'s are continuously differentiable, $\bar{\Theta}$ is the closure of its interior. Let $\pi_{\bar{\Theta}}(y)$ denotes any closet point in $\bar{\Theta}$ to y. Then the projected algorithm is

$$\theta_t = \pi_{\bar{\Theta}}(\theta_{t-1} + \gamma_t B(\theta_{t-1}, \xi_t)). \tag{2.21}$$

Several ordinary differential equations (ODE) methods for proving convergence of θ_t have been developed in the literature. The aim of these methods is to get an ODE, which can be written for 2.20 as

$$\dot{\theta} = \int B(\theta, \xi) P_\theta(d\xi), \tag{2.22}$$

where $P_\theta(\cdot)$ is the stationary distribution of the sequence ξ_t, when $\theta_t \equiv \theta$.

Below we consider a case where ξ_j is not state-dependent and satisfies a type of ϕ-mixing condition. ξ_j is not state-dependent; i.e., for all j

$$P(d\xi_j|\xi_{u-1}, \theta_{u-1}, u \leq j) = P(d\xi_j|\xi_{u-1}, u \leq j).$$

Define a Markov process $\xi_n, n \geq 0$ on S via the transition function $P(\xi, l, \cdot)$, where $P(\xi, l, B) = \int P(\xi, l-k, d\xi')P(\xi', k, B)$ is defined recursively, starting with the given $P(\xi, 1, d\xi')$.

The convergence of the SA algorithm can be analyzed under the following conditions:

Assumption 2.4.1. *$\theta_t \in \Theta$, a compact subset of $\mathbb{R}^{d_\theta}$; and $\xi_t \in S$, a compact subset of $\mathbb{R}^{d_\xi}$.*

Assumption 2.4.2. *$B(\cdot, \cdot)$ is bounded.*

Assumption 2.4.3. *For the Markov process with transition function $P(\xi, j, \cdot)$, there is a unique invariant measure $P(\cdot)$.*

Assumption 2.4.4. *The transition $P(\xi, 1, \cdot)$ is weakly continuous in ξ and such that for each bounded and continuous function $g(\cdot)$, with $G(\xi) = \int P(\xi, i, d\xi)g(\xi)$,*

$$\lim_{j-t\to\infty, t\to\infty} = \left[\int P(d\xi_i|\xi_{j-1}; \xi_{u-1}, u \leq t)g(\xi_j) - G(\xi_{j-1})\right] = 0, \ \ a.s.$$

Assumption 2.4.5. *Define $F(\theta, \xi) = intB(\theta, \xi')P(\xi, 1, d\xi')$. Then $F(\cdot, \cdot)$ is continuous and*

$$\lim_{j-t\to\infty, t\to\infty} = \left[\int P(d\xi_i|\xi_{j-1}; \xi_{u-1}, u \leq t)B(\theta_{j-1}, \xi_j) - F(\theta_{j-1}, \xi_{j-1})\right] = 0, \ \ a.s.$$

Assumption 2.4.6. $\sum_t |\gamma_t - \gamma_t| \leq \infty, \quad \gamma_t > 0, \quad \gamma_t \to 0, \quad \sum_t \gamma_t = \infty.$

Lemma 2.4.3. *(Kushner and Shwartz (1984), p.22) Assume (2.4.1)-(2.4.6) hold. Then* $\theta_t(\cdot)$ *is tight and the limit* $\theta(\cdot)$ *of any weakly convergent subsequence satisfies*

$$\dot{\theta} = \int B(\theta, \xi) P(d\xi), \quad a.s., \tag{2.23}$$

where $\theta(0) \in \Theta$*, and the right-hand side is continuous in* θ*.*

If (2.21) is used in lieu of (2.20) and θ_t is in some Euclidean space, then $\theta_t(\cdot)$ is tight and the limit $\theta_t(\cdot)$ of any weakly convergent subsequence satisfies the projected equation

$$\dot{\theta} = \bar{\pi}(\int B(\theta, \xi) P(d\xi)), \quad \text{a.s.}, \tag{2.24}$$

where $\bar{\pi}(h(\dot{}))$ denotes the projection of the vector field $h(\cdot)$ onto $\bar{\Theta}$.

Lemma 2.4.4. *Kushner and Shwartz (1984), p.23) Let (2.20) be replaced by the following equation:*

$$\theta_t = \theta_{t-1} + \gamma_t B(\omega_t), \tag{2.25}$$

ω *denotes a vector of random variables. Suppose that there is a continuous* $F(\cdot)$ *such that*

$$E[B(\omega_{j+1}|\theta_j, \xi_j; \theta_u, \xi_u, u \leq t)] - F(\theta_j, \xi_j) \to 0, \text{ in probability,} \tag{2.26}$$

as $j - t \to \infty$ *and* $t \to \infty$*. Then Lemma (2.4.3) continues to hold.*

2.4.3 Convergence of the proposed ensemble Kalman filter

Let $\mathbf{y}_{1:t} = (\mathbf{y}_1, \ldots, \mathbf{y}_t)$ denote the observation vector up to time t, and let $\mathbf{x}_{1:t-1} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$ denote the state vector up to time $t-1$. Naturally, $\mathbf{x}_{1:t-1}$ can be treated as missing values of the system, and the complete data likelihood function is given by

$$p(\mathbf{y}_{1:t}, \mathbf{x}_{1:t-1}|\theta) = \prod_{i=1}^{t} p(\mathbf{y}_i|\mathbf{x}_{i-1}, \theta) p(\mathbf{x}_{t-1}), \tag{2.27}$$

where $p(\mathbf{x}_{1:t-1})$ denotes the joint density function of $x_{1:t-1}$. Under regularity conditions, the MLE of θ can be found by solving the system of equations:

$$\begin{aligned} E_{x|\theta}\left[\frac{\partial}{\partial\theta}\log p(\mathbf{y}_{1:t},\mathbf{x}_{1:t-1}|\theta)\right] &= E_{x|\theta}\left[\sum_{i=1}^{t-1}\frac{\partial}{\partial\theta}\log p(\mathbf{y}_{i+1}|\mathbf{x}_i|\theta)\right] \\ &= \frac{\partial}{\partial\theta}\log p(\mathbf{y}_{1:t}|\theta)=0, \end{aligned} \tag{2.28}$$

where $E_{x|\theta}[\cdot]$ denotes the expectation with respect to the density $p(\mathbf{x}_{1:t-1}|\theta)$. If we further assume that the dynamic system (2.17) is stationary, then (2.28) can be simplified as

$$h(\theta)E_{x,y|\theta}\left[\frac{\partial}{\partial\theta}\log p(y_t|x_{t-1},\theta)\right]=0, \tag{2.29}$$

where $E_{x,y|\theta}[\cdot]$ denotes the expectation with respect to the joint density $p(x_{t-1},y_t|\theta)$.

On the convergence of the proposed EnKF algorithm with parameter estimation, we establish the following two theorems. Theorem 2.4.1 concerns the convergence of the proposed EnKF algorithm for the linear system, i.e., $f(\mathbf{x}_t,\alpha)=F(\alpha)\mathbf{x}_t$ for some function $F(\alpha)$.

Theorem 2.4.1. *For a linear system with additive Gaussian white noises and Gaussian initial condition, if the following conditions are satisfied:*

(i) the system is stationary, the observation sequence $\mathbf{y}_t$ *satisfies the conditions (2.4.3)-(2.4.5), and* $\mathbf{y}_t$ *can be included in a compact set,*

(ii) the gain factor sequence γ_t *satisfies the condition (2.4.6),*

(iii) the function $\phi(\mathbf{y},\mathbf{x},\theta)$ *is a continuous function of* $(\mathbf{y},\mathbf{x},\theta)$ *and also a Lip(C,*κ*) function of* $\mathbf{x}$ *for some constants* C *and* κ*,*

(iv) the initial ensemble $\mathbf{X}_0^f$ *has finite moments of any order;*

then $\theta_t \to \theta^0$ *as* $t \to \infty$ *and the emsemble size* $m \to 1$*, where* θ^0 *denotes a solution to (2.29).*

Proof. For the EnKF algorithm, it follows from (12) that $\mathbf{y}_t$ corresponds to ξ_t in the stochastic approximation algorithm (2.20). Since $\mathbf{y}_t$ can be included in a compact set, S is compact. Further, by restricting Θ to be a compact set, this verifies condition (2.4.1). Since $\phi(\mathbf{y}, \mathbf{x}, \theta)$ is a continuous function of $(\mathbf{y}, \mathbf{x}, \theta)$, $E_x\phi(\mathbf{y}, \mathbf{x}, \theta)$ is continuous in both $\mathbf{y}$ and θ and thus bounded on $S \times \Theta$, where E_x denotes the expectation with respect to the Bayesian filter distribution μ, i.e., $p(\mathbf{x}_t|\theta, \mathbf{y}_{1:k})$ for the ensemble $\mathbf{x}_t^a$. This verifies condition (2.4.2). The conditions (2.4.3)-(2.4.6) are satisfied by the assumptions of this theorem. This theorem can then be concluded following from (2.24), where ω denotes the ensemble produced by the EnKF algorithm. Note that since the system is linear, by Lemma (2.4.1) and Lemma (2.4.2), (2.26) holds as $m \to \infty$. □

Theorem 2.4.2 concerns the convergence of the proposed EnKF algorithm for the nonlinear system, where $f(\mathbf{x}_t, \alpha)$ is a nonlinear function of $\mathbf{x}_t$.

Theorem 2.4.2. *For a nonlinear system with additive Gaussian white noises and Gaussian initial condition, if the following conditions are satisfied:*

(i) the system is stationary, the observation sequence $\mathbf{y}_t$ *satisfies the conditions (2.4.3)-(2.4.5), and* $\mathbf{y}_t$ *can be included in a compact set,*

(ii) the gain factor sequence γ_t *satisfies the condition (2.4.6),*

(iii) the function $\phi(\mathbf{y}, \mathbf{x}, \theta)$ *is a continuous function of* $(\mathbf{y}, \mathbf{x}, \theta)$ *and also a Lip(C,*κ*) function of* $\mathbf{x}$ *for some constants* C *and* κ*.*

(iv) the initial ensemble $\mathbf{X}_0^f$ *has finite moments of any order,*

(v) $\bar{h}(\theta_t) = h(\theta_t)$ *for all* $k \geq 0$, *where* $\bar{h}(\theta_t) = \int \phi(\mathbf{y}_t, \mathbf{x}^a_{t-1}, \theta_t)\bar{\mu}^a_t(d\mathbf{x}^a_{t-1})$ *is the mean of* $\phi(\cdot)$ *with respect to the decoupled EnKF distribution* $\bar{\mu}^a_t$, *and* $h(\theta_t) = \int \phi(\mathbf{y}_t, \mathbf{x}^a_{t-1}, \theta_t)$ $\mu_t(d\mathbf{x}^a_{t-1})$ *is the mean of* $\phi(\cdot)$ *with respect to the Bayesian filter distribution* μ_t,

then $\theta_t \to \theta^0$ *as* $t \to \infty$ *and the emsemble size* $m \to 1$, *where* θ^0 *denotes a solution to (2.29).*

Proof. It follows from the proof of Theorem 2.4.1 directly. If condition (v) is satisfied, then it follows from Lemma (2.4.1) and Lemma (2.4.2) that (2.26) can still hold. □

2.5 Simulation Study

In this section I used multiple simulation studies to evaluate the performance of our proposed method. We have compared our performance to that of augmentation based parameter estimation technique. In the following studies, we have showed that our method can successfully estimate various model parameters. The simulation study is divided into two sub sections. In the first section, we have considered linear spatio-temporal models. In the second sub-section, we looked at 40-variable non-linear system of Lorenz (1996).

2.5.1 Spatio-temporal Models

In this section, I have done simulation study based on spatio-temporal models. I started with basic auto regressive model. Gradually, I progressed to more complex spatio-temporal models, and assessed performance of our technique. In most of the simulation exercise the sequence $\{\gamma_t\}$ is of the form

$$\gamma_t = d_0 \left(\frac{n_0}{\max(n_0, t)} \right)^{\nu}, \tag{2.30}$$

where, unless otherwise mentioned, we take $d_0 = 0.01$, $\nu = 1$ and $n_0 = 500$.

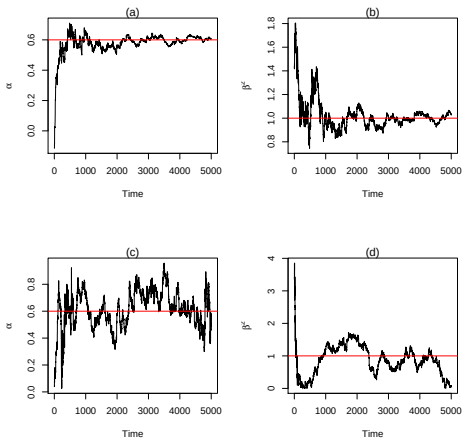

Figure 2.2: The figure shows estimated value of α and β^2 from : (a)-(b) Stochastic Approximation approach; (c)-(d) Augmentation approach. In each of the four cases the horizontal line represents the true value of the corresponding parameter. We have used a time-series of length 5000.

2.5.1.1 Auto Regressive Model of Order 1

Let us consider an AR(1) model

$$x_t = \alpha x_{t-1} + \beta w_t, \tag{2.31}$$

where w_t is Gaussian white noise with zero mean and unit variance. For our model, we take $\alpha = 0.6$ and $\beta = 1$. The observations, y_t, are defined as the true state plus a random error, v_t, where $v_t \sim \mathrm{N}(0, 0.01)$. Corresponding log-likelihood function is given by

$$\log(p(y_t|x_{t-1}, \theta)) \propto -\frac{1}{2}\log(0.01 + \beta^2) - \frac{(y_t - \alpha x_{t-1})^2}{2(0.01 + \beta^2)}.$$

The respective elements of $\phi(y_t, x_{t-1}, \theta)$ are given by

$$\begin{aligned}\phi(y_t, x_{t-1}, \alpha) &= \frac{(y_t - \alpha x_{t-1})x_{t-1}}{(0.01 + \beta^2)}, \\ \phi(y_t, x_{t-1}, \beta) &= -\frac{\beta}{(0.01 + \beta^2)} + \frac{\beta(y_t - \alpha x_{t-1})^2}{(0.01 + \beta^2)^2}.\end{aligned}$$

	State RMSE	α	β^2
SA	0.1002(0.00002)	0.6080(0.0003)	1.0045(0.0005)
Augmentation	0.1236(0.00186)	0.6311(0.1062)	1.2225(0.9670)

Table 2.1: Results from joint State and Parameter estimation of an AR(1) model. True value of α is 0.6 and that of β^2 is 1.0. Results are based on average of 10 independent simulations. The numbers in the parentheses denote the standard error of the estimates (same convention is followed for other tables as well).

We jointly estimate the deterministic parameter α and the stochastic parameter β based on 5000 observations. Initial value of α is randomly drawn from N(0.1,0.05) and that of β is drawn from N(1.5,1). After some initial run, we decided to use an ensemble of size 25. For a typical run, figure (2.2)(a)-(b) show the estimation result of the model parameters using stochastic approximation based approach. Figure (2.2)(c)-(d) show the same results from augmentation based approach. Figure(2.3) shows the state estimation result. The original series in thinned to highlight the performance difference. Every 100 time points is considered. Using trajectory averaging, the mean estimate of α is 0.6076 and that of β is 1.0054. These results are based on last 1000 iteration. For the augmentation based approach, mean of α is 0.5798 and that of β is 0.6476. Table (2.1) gives the numerical comparison with the augmentation based approach based on 10 independent simulations. Results clearly show that the stochastic approximation based technique performs better both in terms of parameter estimation and State estimation than the augmentation based approach.

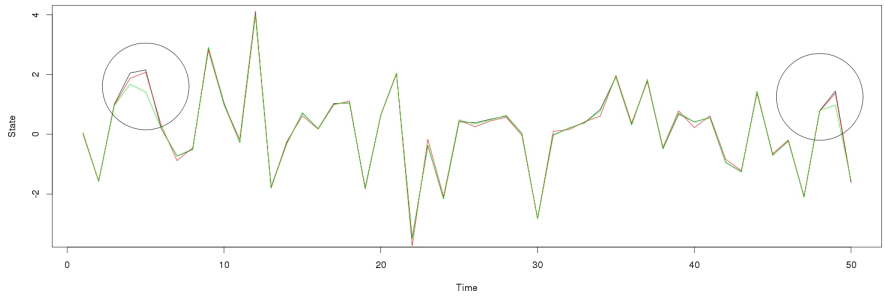

Figure 2.3: Time series plot of True states(in red), states estimated using SA(in black) and states estimated using Augmentation(in green) based approach. Circled areas demonstrate the gain in state estimation from the SA based approach.

2.5.1.2 Multivariate Auto Regressive Model with Spatial Error

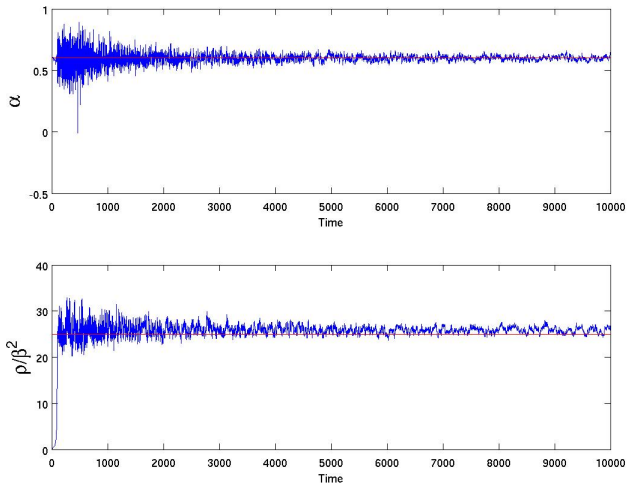

Figure 2.4: The figure shows the estimated value of α and $\frac{\rho}{\beta^2}$ using SA based approach. The horizontal line represents the true value of the corresponding parameter. We have used a time-series of length 10000.

Next we considered the following multi-variate auto regressive model

$$\mathbf{x}_t = \alpha \mathbf{x}_{t-1} + \beta \mathbf{w}_t, \tag{2.32}$$

where $\mathbf{x}_t \in \mathbb{R}^p$ is the p-dimensional state variable, $\mathbf{w}_t = (w_t(s_1), \ldots, w_t(s_p))$ is a zero mean Gaussian spatial process with exponential covariance $\mathbf{R}_\rho = \exp(-\frac{\mathbf{D}}{\rho})$, where $\mathbf{D} = \{||s_i - s_j||\}$ is the distance matrix. Observation $\mathbf{y}_t$ is generated by adding a gaussian random error $\mathbf{v}_t \sim \mathrm{N}(0, 0.01 * \mathbf{I}_p)$ to the underlying state $\mathbf{x}_t$. For our simulation, model is generated using $\alpha = 0.6, \beta = 1$ and $\rho = 25$. The log-likelihood function can be written as

$$\log(p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta)) \propto -\frac{p}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1}), \tag{2.33}$$

where $\mathbf{\Sigma} = 0.01 \times \mathbf{I}_p + \beta^2\mathbf{R}_\rho$. Differentiating (2.33) with respect to θ, we get

$$\begin{aligned}
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \alpha) &= (\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}\mathbf{x}_{t-1}, \\
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \beta) &= -\beta[\mathrm{trace}(\mathbf{\Sigma}^{-1}\mathbf{R}_\rho) - (\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}\mathbf{R}_\rho\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})], \\
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \rho) &= -\left(\frac{\beta^2}{2\rho^2}\right)[\mathrm{trace}(\mathbf{\Sigma}^{-1}(\mathbf{D} \cdot \mathbf{R}_\rho)) \\
&\quad -(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}(\mathbf{D} \cdot \mathbf{R}_\rho)\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})],
\end{aligned}$$

where $\mathbf{A} \cdot \mathbf{B}$ denote the element wise matrix multiplication.

Using the package geoR (Jr and Diggle (2001)), we simulated a data set of size T = 10,000 with $p = 100$ sampling sites uniformly distributed in a bounded region of $[0, 100] \times [0, 100]$. Multiple initial runs indicated that an ensemble of size 125 is sufficient to successfully estimate model states and the parameter vector. Initial values of α, β and ρ are drawn from N(0.2,0.05), N(1,1.5) and N(25,3)respectively. In case of ρ, we used $d_0 = 1$ to calculate γ_t.

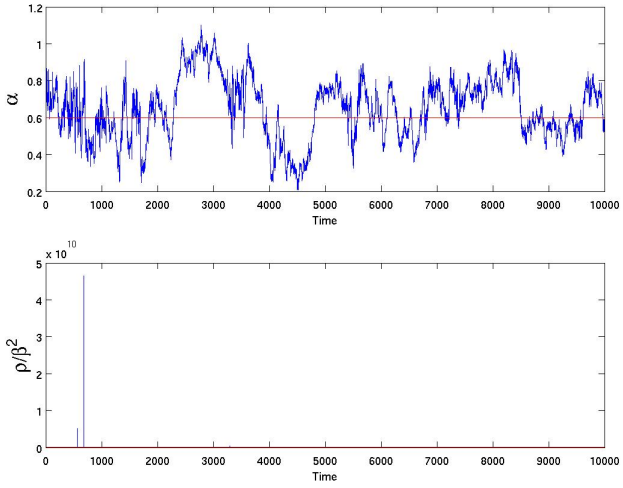

Figure 2.5: The figure shows the estimated value of α and $\frac{\rho}{\beta^2}$ from Augmentation based approach. The horizontal line represents the true value of the corresponding parameter. We have used a time-series of length 10000.

Numerical results, based based on the trajectory mean of last 2000 iterations, averaged over 10 independent runs, are presented in table (2.2). Figure (2.4) shows the SA based estimation result from a typical run. Similar result from a augmentation based approach can be found in figure (2.5). Clearly augmentation based approach fails to estimate the model parameters. In case of exponential correlation function, Stein (2004),Stein (1999) show that model (2.32) suffers from identifiability problem. For $\frac{\rho_1}{\beta_1^2} = \frac{\rho_2}{\beta_2^2}$, probability measures can be identical for $w_t(s), s \in A$, for any bounded A $\in \mathbb{R}^p$. Hence in table (2.2), we reported ratio of β^2 and ρ. Figure (2.6) shows the mean square error in state estimation over time. For state estimation, SA based approach produces much stable results than the augmentation based approach.

2.5.1.3 Random Coefficient Autoregressive Model with Spatial Error

Next we add a random coefficient to the auto regressive part in equation (2.32). Random coefficient autoregressive(RCA) model was introduced in Nicholls and Quinn

	State RMSE	α	β^2	Ratio
SA	0.1316(0.0055)	0.5969(0.0003)	1.4045(0.2876)	24.1581(0.3227)
Augmentation	0.1397(0.0062)	0.6008(0.1089)	5.3213(1.3752)	8.2775(1.8472)

Table 2.2: Numerical results from Multivariate Auto-regressive model with spatial error. True value of α, β^2 and Ratio($\frac{\rho}{\beta^2}$) are 0.6, 1 and 25 respectively. Results are based on average of 10 independent simulations.

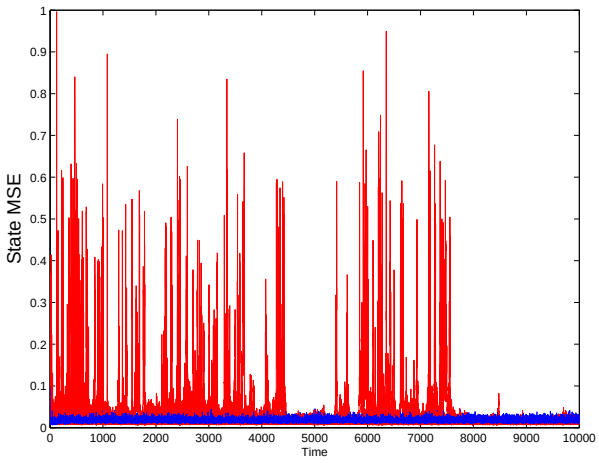

Figure 2.6: Time series plots of Mean Square Error from state estimation: Red line is the State MSE from SA based approach and the Blue line is the same from Augmentation based approach. MSE from all 10 runs are plotted.

(1982). Later Aue et al. (2006), Aue and Horváth (2011) looked at various properties of the RCA model. Consider the the following model

$$\mathbf{x}_t = (\alpha + b_t)\mathbf{x}_{t-1} + \beta \mathbf{w}_t, \tag{2.34}$$

where $\mathbf{x}_t \in \mathbb{R}^p$ is the p-dimensional state variable, $b_t \sim \mathrm{N}(0, \sigma^2)$ is the random coefficient, $\mathbf{w}_t = (w_t(s_1), \ldots, w_t(s_p))$ is a zero mean Gaussian spatial process with exponential covariance $\mathbf{R}_\rho = \exp(-\frac{\mathbf{D}}{\rho})$, where $\mathbf{D} = \{||s_i - s_j||\}$ is the distance matrix. $\{(b_t, \mathbf{w}_t) : t \in \mathbb{Z}\}$ are independently distributed. Under some suitable condition, $\mathbf{x}_t$ is

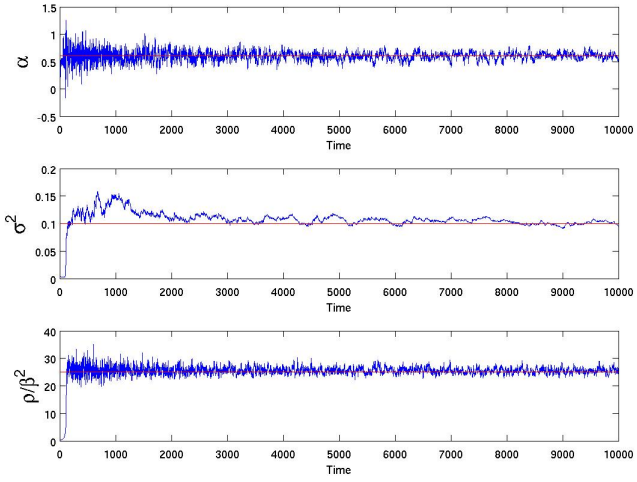

Figure 2.7: Estimated values of α, σ^2 and $\frac{\rho}{\beta^2}$ using SA based approach. The horizontal line represents the true value of the corresponding parameter. We have used a time-series of length 10000.

stationary iff $(\alpha^2 + \sigma^2 < 1)$. We can re-write equation (2.34) as

$$\mathbf{x}_t = \alpha \mathbf{x}_{t-1} + \mathbf{e}_t, \tag{2.35}$$

where $\mathbf{e}_t|\mathbf{x}_{t-1} \sim \mathrm{N}(0, \sigma^2(\mathbf{x}_{t-1}\mathbf{x}'_{t-1}) + \beta^2\mathbf{R}_\rho)$. Observation $\mathbf{y}_t$ is generated by adding a gaussian random error $\mathbf{v}_t \sim \mathrm{N}(0, 0.01 * \mathbf{I}_p)$ to the underlying state $\mathbf{x}_t$. The log-likelihood function is given by

$$\log(p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta)) \propto -\frac{p}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1}), \tag{2.36}$$

where $\mathbf{\Sigma} = 0.01 \times \mathbf{I}_p + \sigma^2(\mathbf{x}_{t-1}\mathbf{x}'_{t-1}) + \beta^2\mathbf{R}_\rho = 0.01 \times \mathbf{I}_p + \sigma^2\mathbf{U}_{t-1} + \beta^2\mathbf{R}_\rho$. Corresponding $\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \theta)$ is given by

$$
\begin{aligned}
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \alpha) &= (\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}\mathbf{x}_{t-1}, \\
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \sigma) &= -\sigma[\text{trace}(\mathbf{\Sigma}^{-1}\mathbf{U}_{t-1}) - (\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}\mathbf{U}_{t-1}\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})], \\
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \beta) &= -\beta[\text{trace}(\mathbf{\Sigma}^{-1}\mathbf{R}_\rho) - (\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}\mathbf{R}_\rho\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})], \\
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \rho) &= -\left(\frac{\beta^2}{2\rho^2}\right)[\text{trace}(\mathbf{\Sigma}^{-1}(\mathbf{D} \cdot \mathbf{R}_\rho)) \\
&\quad -(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}(\mathbf{D} \cdot \mathbf{R}_\rho)\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})].
\end{aligned}
$$

For our simulation, model is generated using $\alpha = 0.6, \sigma^2 = 0.1, \beta^2 = 1$ and $\rho = 25$.

As before I used geoR to simulate a data set for 10,000 time points with $p = 100$ sampling sites uniformly distributed in a bounded region of $[0, 100] \times [0, 100]$. Initial values of α, σ, β and ρ are generated using N(0.2,0.05), N(0.2,0.1), N(1,1.5) and N(25,3). In (2.30), we take $\nu = 0.7$. Also, for α, ρ and σ, we use d_0 as 0.05, 1 and 0.001 respectively. After several initial runs we decided to use an ensemble of size 125. Due to the identifiability problem mentioned before, we looked at the ratio ρ/β^2 rather than individual estimates. Table (2.3) shows the result, based on the trajectory average of last 2000 iterations, averaged over 10 independent runs. Figure (2.7) and (2.8) show the SA and augmentation based result from a typical run. Clearly augmentation based approach fails to estimate the stochastic parameters. MSE over time, for all 10 independent runs, are plotted in Figure (2.9). We notice that augmentation based approach fail to estimate unobserved state for the initial data, although they perform as well for the long run. In SA based approach, MSE is stable throughout the series.

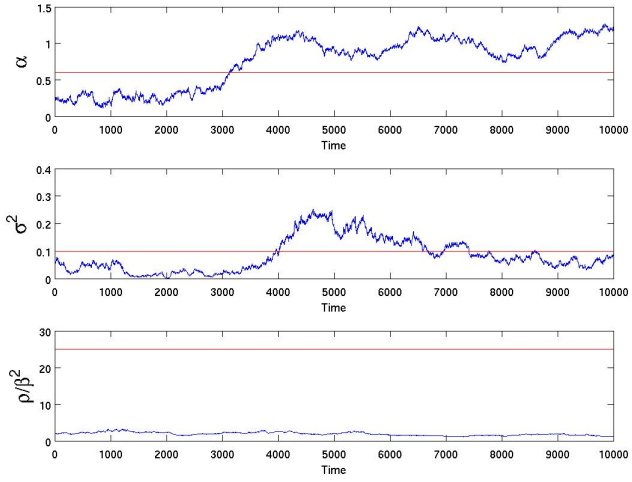

Figure 2.8: Parameter estimation for the Random Coefficient Auto Regressive Model with Spatial Error using augmentation based approach. The horizontal line represents the true value of the corresponding parameter. We have used a time-series of length 10000.

	State RMSE	α	σ^2	Ratio
SA	0.1315(0.0055)	0.6102(0.0008)	0.09997(0.0002)	23.5414(0.6902)
Augmentation	0.1231(0.0055)	0.5882(0.0475)	3.2467(2.3352)	15.0779(13.2125)

Table 2.3: True value of α, σ^2 and Ratio($\frac{\rho}{\beta^2}$) are 0.6, 0.1 and 25 respectively. Results are based on average of 10 independent simulations.

2.5.1.4 Spatial Random Coefficient Autoregressive Model with Spatial Error

Finally we considered a spatial random Coefficient model with spatial error given by

$$x_t(s) = (\alpha + b_t(s))x_{t-1} + \beta w_t(s), \tag{2.37}$$

where $\mathbf{x}_t = [x_t(s_1), \ldots, x_t(s_p)]$ is the p-dimensional state vector, $\mathbf{b}_t = [b_t(s_1), \ldots, b_t(s_p)]$ $\sim \mathrm{N}(0, \sigma^2\mathbf{R}_\delta)$ is the spatial random coefficient and $\mathbf{w}_t$ is the Gaussian error as described in equation (2.34). Here $\mathbf{R}_\delta = \exp(-\frac{\mathbf{D}}{\delta})$ is the exponential covariance function where $\mathbf{D}$

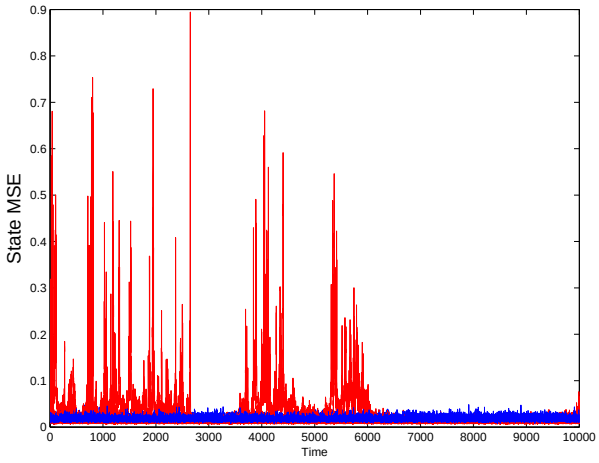

Figure 2.9: Time series plot of Mean Square Error from state estimation for all 10 independent runs: Red line is the State MSE from SA based approach and the Blue line is the same from Augmentation based approach.

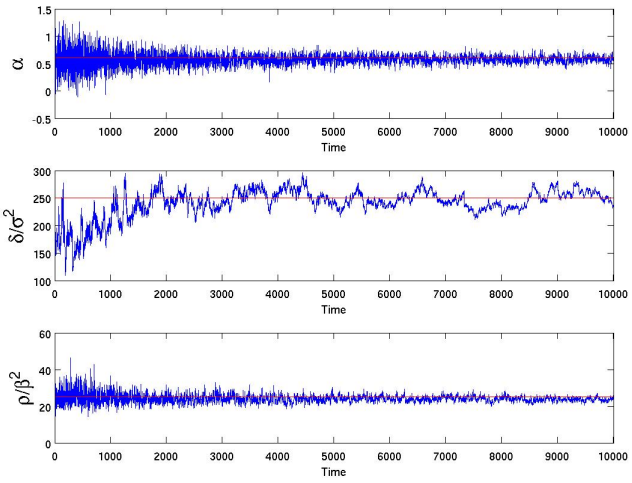

Figure 2.10: Estimated values of α, $\frac{\delta}{\sigma^2}$ and $\frac{\rho}{\beta^2}$ using Stochastic Approximation. The horizontal line represents the true value of the corresponding parameter. We have used a time-series of length 10000.

is the distance matrix. Model assumptions are same as before. As before, we can re-write

(2.37) as

$$\mathbf{x}_t = \alpha \mathbf{x}_{t-1} + \mathbf{u}_t, \tag{2.38}$$

where $\mathbf{u}_t|\mathbf{x}_{t-1} \sim \mathrm{N}(0, \sigma^2 \mathbf{X}_{t-1}\mathbf{R}_\rho \mathbf{X}_{t-1} + \beta^2 \mathbf{R}_\rho)$, $\mathbf{X}_{t-1} = \mathrm{diag}(x_t(s_1), \ldots, x_t(s_p))$. By adding a gaussian random error $\mathbf{v}_t \sim \mathrm{N}(0, 0.01 * \mathbf{I}_p)$ to the state $\mathbf{x}_t$, observation $\mathbf{y}_t$ is generated. The log-likelihood function for (2.38)

$$\log(p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta)) \propto -\frac{p}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1}), \tag{2.39}$$

where $\mathbf{\Sigma} = 0.01 \times \mathbf{I}_p + \sigma^2 \mathbf{X}_{t-1}\mathbf{R}_\delta \mathbf{X}_{t-1} + \beta^2 \mathbf{R}_\rho$.

We get $\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \theta)$ by differentiating (2.38) w.r.t θ as,

$$\begin{aligned}
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \alpha) &= (\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}\mathbf{x}_{t-1}, \\
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \sigma) &= -\sigma[\mathrm{trace}(\mathbf{\Sigma}^{-1}\mathbf{X}_{t-1}\mathbf{R}_\delta\mathbf{X}_{t-1}) \\
&= -(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}(\mathbf{X}_{t-1}\mathbf{R}_\delta\mathbf{X}_{t-1})\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})], \\
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \delta) &= -\left(\frac{\sigma^2}{2\delta^2}\right)[\mathrm{trace}(\mathbf{\Sigma}^{-1}\mathbf{X}_{t-1}(\mathbf{D}\cdot\mathbf{R}_\delta)\mathbf{X}_{t-1}) \\
&\quad -(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}\mathbf{X}_{t-1}(\mathbf{D}\cdot\mathbf{R}_\delta)\mathbf{X}_{t-1}\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})] \\
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \beta) &= -\beta[\mathrm{trace}(\mathbf{\Sigma}^{-1}\mathbf{R}_\rho) - (\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}\mathbf{R}_\rho\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})], \\
\phi(\mathbf{y}_t, \mathbf{x}_{t-1}, \rho) &= -\left(\frac{\beta^2}{2\rho^2}\right)[\mathrm{trace}(\mathbf{\Sigma}^{-1}(\mathbf{D}\cdot\mathbf{R}_\rho)) \\
&\quad -(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})'\mathbf{\Sigma}^{-1}(\mathbf{D}\cdot\mathbf{R}_\rho)\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \alpha\mathbf{x}_{t-1})].
\end{aligned}$$

Parameters used for simulation purpose are $\alpha = 0.6, \sigma^2 = 0.1, \delta = 25, \beta^2 = 1$ and $\rho = 25$.

We have simulated a data set for 10,000 time points with $p = 50$ sampling sites uniformly distributed in a bounded region of $[0, 100]\times[0, 100]$. Initial values of $\alpha, \sigma, \delta, \beta$ and ρ are generated using N(0.2,0.05), N(0.2,0.1), N(25,3), N(1,1.5) and N(25,3). In (2.30), for α, σ, δ and ρ, we use d_0 as 0.05, 0.002, 5 and 1 respectively. After several initial runs we decided to use an ensemble of size 100. Due to the identifiability problem mentioned

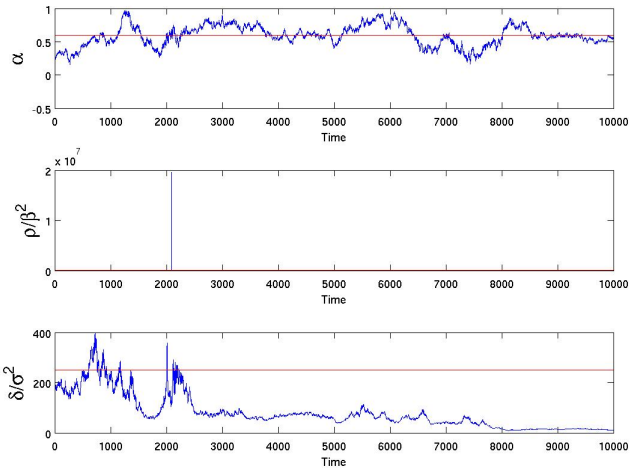

Figure 2.11: Estimated values of α, $\frac{\delta}{\sigma^2}$ and $\frac{\rho}{\beta^2}$ using Augmentation based approach. The horizontal line represents the true value of the corresponding parameter.

before, we looked at the ratios δ/σ^2 and ρ/β^2 rather than individual estimates. Trajectory averaging based numerical results, based on the of last 2000 iterations, averaged over 10 independent runs, are shown in table (2.4). Figure (2.10) and (2.11) show the SA and augmentation based result from a typical run. As models become more complicated, augmentation based approach fail to estimate model parameters. Figure (2.12) shows the MSE over time for both approaches. MSE from all 10 independent runs are plotted here. The same criticisms as before are valid in this case as well.

	State RMSE	α	$\frac{\delta}{\sigma^2}$	$\frac{\rho}{\beta^2}$
SA	0.1042(0.0034)	0.5758(0.0002)	254.8377(2.6589)	23.7038(0.0241)
Augmentation	0.1044(0.0034)	0.6195(0.0617)	133.3108(50.4696)	5.6881(1.0574)

Table 2.4: True value of α, $\frac{\delta}{\sigma^2}$ and $\frac{\rho}{\beta^2}$ are 0.6, 250 and 25 respectively. Results are based on average of 10 independent simulations.

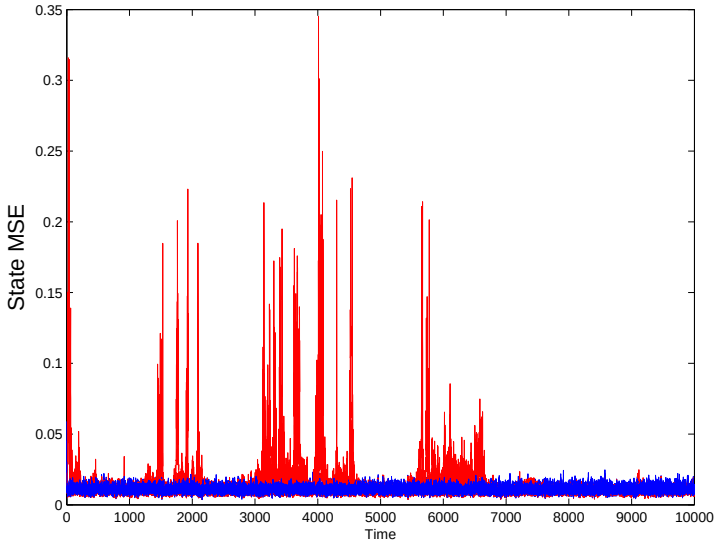

Figure 2.12: Time series plot of Mean Square Error from state estimation: Red line is the State MSE from SA based approach and the Blue line is the same from Augmentation based approach.

2.5.2 Lorenz-96 Model

The model used here is an extended version of Lorenz-96 model, introduced by Lorenz and Emanuel (1998). This extended version of the model is similar to the model described in Yang and Delsole (2009). The model here is given by

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - \frac{x_i}{1+d_i} + f_0 + f_i, \tag{2.40}$$

where $i = 1, 2, \ldots, 40$, f_0 is 8.0, f_i's and d_i's are randomly chosen from N$(0, 1)$ and $N(0.25, 0.2)$ respectively. We reject observed d_i's if the value is less than or equal to -1. Differential equation in (2.40) is solved using fourth-order Runge-Kutta numerical method. Time interval used for integrating the model forward is 0.05. We get the observed data by adding standard Gaussian random noise to the true states. Here data is observed at half of the randomly selected grid points. The observation model is given

by

$$\mathbf{y}_t = \mathbf{H}\mathbf{x}_t + \boldsymbol{\epsilon}_t, \tag{2.41}$$

where $\boldsymbol{\epsilon}_t \sim \mathrm{N}(0, \mathbf{I}_q)$. Following Whitaker and Hamill (2002) and Yang and Delsole (2009) we use Ensemble Square Root Filter for state estimation, and then apply our method for parameter estimation. In all the different cases we kept ensemble size to 20 to compare our results to Yang and Delsole (2009).

2.5.2.1 Estimating The Additive Parameter

Next we estimate the additive parameter $\mathbf{f}$. The corresponding $\phi()$ is given by

$$\phi(\mathbf{y}, \mathbf{x}, f_i) = \Delta t(\mathbf{y} - \mathbf{H}\mathbf{x})'\mathbf{H}e_i, \quad i = \ldots, 40,$$

where e_i is a column of zeros with 1 in the i^{th} element. The initial values are randomly drawn from $\mathrm{N}(0, 0.1)$. We ran the system for 36,000 time points. The parameter estimates and the RMSE's are based on last 2,000 time steps. The average root mean square error(RMSE), averaged over 10 independent simulations, for parameter estimation is 0.0653 and that for the state estimation is 0.3476. Figure 2.13 shows observed and estimated $\mathbf{f}$ in one such simulation.

2.5.2.2 Estimating The Multiplicative Parameter

Next we estimate the multiplicative parameter $\mathbf{d}$. Since d_i has to be greater than -1, instead of $\mathbf{d}$, we estimate $z_i = \log(d_i + 1)$. In this case the $\phi()$ function is calculated as

$$\phi(\mathbf{y}, \mathbf{x}, \mathbf{z}) = \Delta t\mathbf{H}\ \mathrm{diag}(\mathbf{x}\exp(-\mathbf{z}) + \frac{\Delta t}{6}(K_1 + K2 + K3)\exp -\mathbf{z})'(\mathbf{y} - \mathbf{H}\mathbf{x}),$$

where $K_1, K_2, K3$ are terms from Runge-Kutta method. The initial values of $\mathbf{d}$ are randomly selected from $\mathrm{N}(0, 0.01)$, rejecting intial values that are less than or equal to 1. We ran the system for 36,000 time points. The parameter estimates and RMSE's are

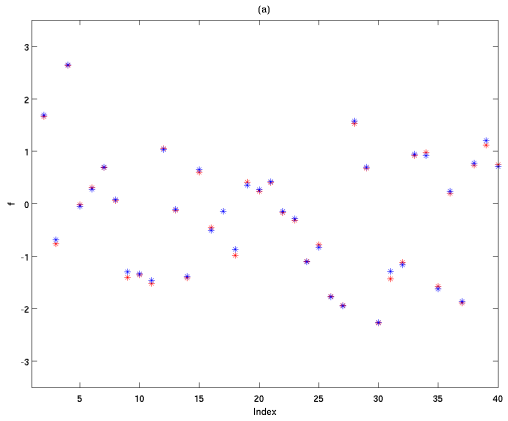

Figure 2.13: Observed(in red) and Estimated(in blue) Additive Parameter f.

based on last 2,000 time points. We have estimated 40 hidden state variables and 40 multiplicative parameters based on 20 observations. RMSE, averaged over 10 independent simulations, for the state estimation is 0.3606 and that for the **d** is 0.0593. Figure 2.14 shows the observed and estimated **d** for one MCMC simulation.

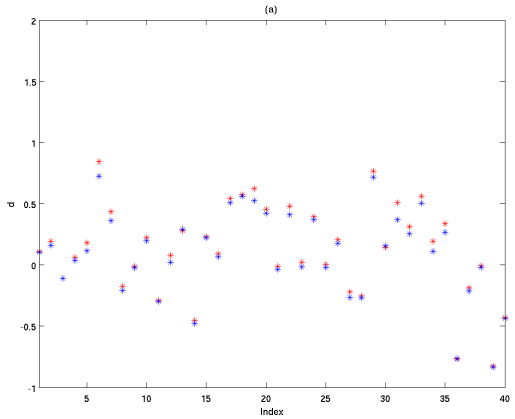

Figure 2.14: Observed(in red) and Estimated(in blue) Multiplicative Parameter d.

2.5.2.3 Joint Estimation of Both the Additive Parameter and the Multiplicative Parameter

In this section we simultaneously estimate both f and d. As before, the initial values of **f** are randomly drawn from $N(0, 0.1)$ and the initial values of **d** are randomly drawn from $N(0, 0.01)$. Again we reject intial values **d** that are less than or equal to 1. We ran the experiment for 36,000 time points. The parameter estimates and calculated RMSE's are based on last 2,000 time steps. Here we have 20 observations, 40 hidden state variable, 40 additive parameters and 40 multiplicative parameters. The RMSE, averaged over 10 different simulation, for state estimation is 0.3682, for **f** is 0.1100 and that of **d** is 0.0685. Figure (2.15) shows the obsrved and estimated **f** and **d** for one such simulation.

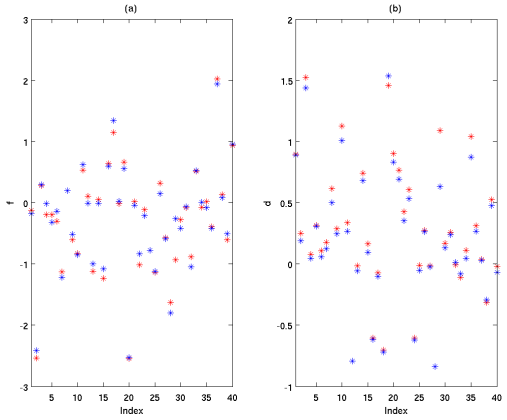

Figure 2.15: Observed(in red) and Estimated(in blue) (a)Additive Parameter f and (b)Multiplicative Parameter d.

2.6 Estimating Parameter Using Stochastic Approximation in Ensemble Kalman Filter - under Limited Data

Consider the dynamic model given in equation (2.17), and we want to solve the problem of parameter estimation using Stochastic Approximation. But unlike before,

	Additive	Multiplicative	Joint
State RMSE	0.3476(0.0062)	0.3606(0.0085)	0.3682(0.0132)
f RMSE	0.0653(0.0137)	-	0.1100(0.0173)
d RMSE	-	0.0593(0.0074)	0.0685(0.0120)
State RMSE(*)	0.3751	0.3775	0.4197

Table 2.5: The Root Mean Square Error(RMSE) for different analysis. The RMSE reported in rows 1-3 are avaraged over 10 independent simulations. Standard errors are reported in the parenthasis. The result in row 4 is taken from Yang and Delsole (2009). That result corresponds to the temporal smoothing parameter $\beta = 0.8$. In all the 3 cases, our method produces better RMSE for the hidden state.

suppose we have limited data. In that case, we repeat the parameter estimation over the same observed data set multiple number of times. The following algorithm summarizes steps to solve the problem of parameter estimation under limited data.

Algorithm 2(Adaptive EnKF with Parameter Estimation for Limited Data)

- Get some initial values for the model state $\mathbf{x}_0$ and the model parameters. Let T be the length of the observed data.

  1. **Forecast Step** Set

$$\begin{aligned} \mathbf{x}_t^{f_i} &= f(\mathbf{x}_{t-1}^{a_i}, \mathbf{u}_{t-1}, \alpha_{(j-1)*T+t-1}) + \mathbf{w}_t^i, \\ \mathbf{y}_t^{f_i} &= \mathbf{H}(\beta_{(j-1)*T+t-1})\mathbf{x}_t^{f_i} + \mathbf{v}_t^i, \end{aligned}$$

  where $\mathbf{w}_t^i \sim \mathrm{N}(0, \mathbf{Q}(\eta_{x,(j-1)*T+t-1}))$ and $\mathbf{v}_t^i \sim \mathrm{N}(0, \mathbf{R}(\eta_{y,(j-1)*T+t-1}))$. Calculate $\hat{\mathbf{P}}_{xx_t}^f$.

  2. **State Update**

$$\begin{aligned} \hat{\mathbf{K}}_t &= \hat{\mathbf{P}}_{xx_t}^f \mathbf{H}(\beta_{t-1})[\mathbf{H}(\beta_{t-1})\hat{\mathbf{P}}_{xx_t}^f \mathbf{H}'(\beta_{t-1}) + \mathbf{R}(\eta_{y,t})]^{-1} \\ \mathbf{x}_t^{a_i} &= \mathbf{x}_t^{f_i} + \hat{\mathbf{K}}_t(\mathbf{y}_t - \mathbf{y}_t^{f_i}). \end{aligned}$$

3. **Parameter Update**

$$\theta_{(j-1)*T+t} = \theta_{(j-1)*T+t-1} + \gamma_t \Phi(\mathbf{y}_t, \mathbf{X}^a_{t-1}, \theta_{(j-1)*T+t-1}).$$

- Repeat step (1)-(3) for j = 1, 2, ..., J.

2.6.1 Random Coefficient Autoregressive Model of with Spatial Error - a Limited Data Simulation Study

In this section we revisit the problem of parameter estimation for Multivariate Autoregressive model with spatial error, but under limited data. From model (2.34) we simulate a data set of 1,000 time points with 100 spatial location uniformly distributed in a bounded region of $[0, 100] \times [0, 100]$, using same simulation parameters. After some initial runs, we decided on an ensemble of size 150. For estimation purpose we use J=10. Also, the sequence $\{\gamma_t^j\}$ is of the form

$$\gamma_t^j = d_0 \left(\frac{n_0}{\max(n_0, (j-1) * T + t)} \right)^{\nu}. \tag{2.42}$$

In this case, take $\nu = 0.7$ and $n_0 = 500$. Value of d_0 changes for different parameters. We take d_0 as 0.05,0.01, 1, 0.001 respectively for α, β, ρ and σ. Figure (2.16) shows the result from a typical run. Estimated values of model parameters, averaged over 10 independent run are :

$$\begin{aligned}
\hat{\alpha} &= 0.59631(0.00008), \\
\hat{\sigma}^2 &= 0.09695(0.00004), \\
\hat{\beta}^2 &= 1.02625(0.00009), \\
\hat{\rho} &= 24.2359(0.0031), \\
\frac{\hat{\rho}}{\hat{\beta}^2} &= 23.6832(0.0041).
\end{aligned}$$

Numbers in parenthesis are the corresponding standard errors.

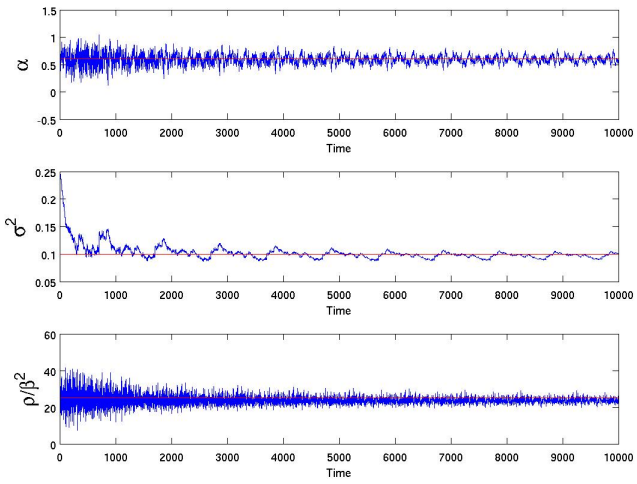

Figure 2.16: Estimated values of α, σ^2 and $\frac{\rho}{\beta^2}$ using SA. The horizontal line represents the true value of the corresponding parameter. We have used a time-series of length 1000, repeated 10 times.

2.7 Real Data Analysis

In this section, we have applied our method to the precipitation data set, obtained from National Climatic Data Center. This data is observed at irregularly spaced 11,918 weather stations, spread all over United States. Precipitation units are in total millimeters per month. This data is collected every month from year 1895 till year 1997. This data set is available at www.image.ucar.edu/GSP/Data/US.monthly.met/. This data set has been part of several studies in the past, e.g., Johns et al. (2003), Furrer et al. (2006), Liang et al. (2013) etc. Johns et al. (2003) imputed parts of the data, but for our purpose, we are going to treat all data as real observation, following Furrer (2006).

For our analysis, we considered 498 weather stations from Colorado. The The total data set is then divided into 2 part - a training data set, comprising of 450 stations,

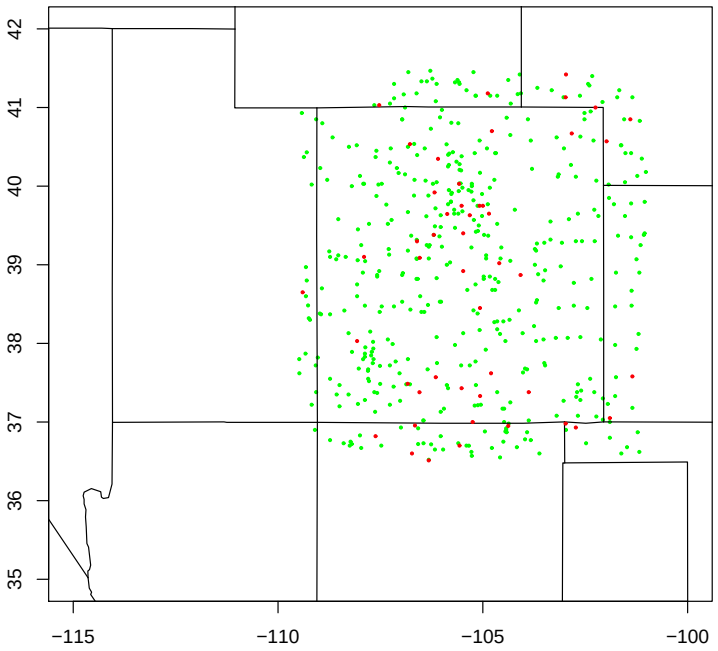

Figure 2.17: Station locations, mostly concentrated in Colorado, in USA map. Green dots represent stations in training data and the red dots are the stations in test data.

and a test data set of 48 stations. Station locations are in figure (2.17). We removed the trend and seasonal component as well as the effect of elevation on temperature from the data using traditional time-series tools. Then we used Stochastic Approximation based parameter estimation method together with Ensemble Kalman filter to estimate model parameters. Due to the size, the training data set was divided into 9 data sets of size 50.The parameter estimates from each data set were later combined using weighted average, inverse standard deviation being the respective weights. Finally we used the estimated model parameters to predict temperature for the test data. We reparametrized the exponential covariance function in equations (2.32), (2.34) and (2.37) to $\mathbf{R}_\rho = \exp(-\rho\mathbf{D})$ and $\mathbf{R}_\delta = \exp(-\delta\mathbf{D})$. Table (2.6) summarizes the numerical results. The Goodness of

Fit Ratio(GoFR) in the table is defined as follows: $1 - \frac{\text{MSE}}{\text{Var(data)}}$.

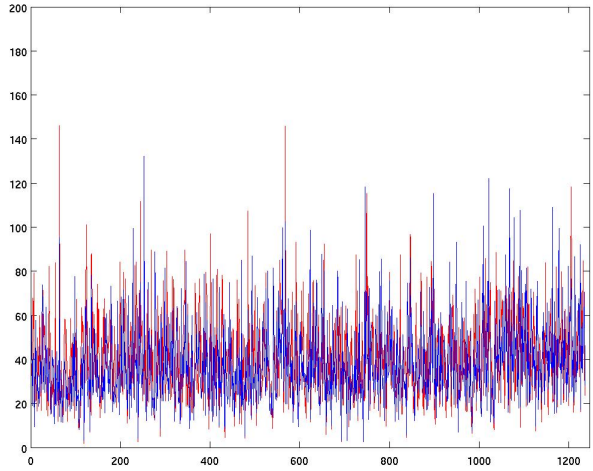

Figure 2.18: True vs. predicted temperature using multivariate autoregression model.

	α	σ^2	δ	β^2	ρ	RMSE	GoFR
MAR	0.268	-	-	1.0372	42.2717	22.5089	64.01%
RCA	0.2859	0.2539	-	0.8542	48.3468	22.4459	64.21%
Spatial RCA	0.2777	0.0302	49.8646	1.0198	47.9083	22.4987	64.04%

Table 2.6: Estimated parameters and the Prediction Root Mean Square Error.

Figure (2.18) show the true and the predicted temperature, averaged over all stations in the test data, for the multivariate autoregression model with spatial error. The red line denote the true temperature, and the blue line is the estimated temperature. Similar figure for the random coefficient model and the spatial random coefficient model can be found in figure (2.19) and (2.20) respectively.

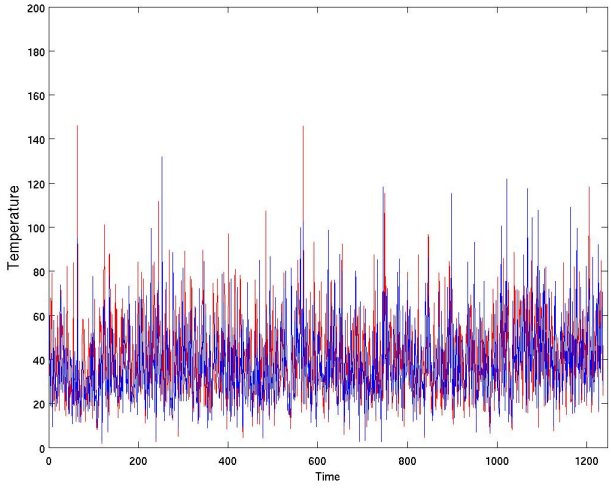

Figure 2.19: True vs. predicted temperature using random coefficient model.

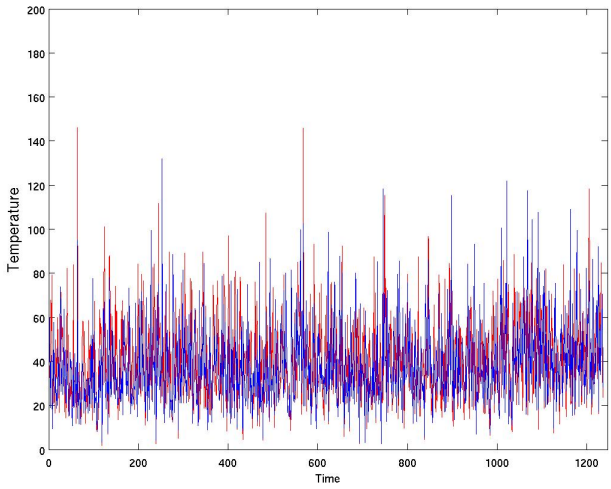

Figure 2.20: True vs. predicted temperature using spatial random coefficient model.

2.8 Conclusion

In our research, we proposed a Stochastic Approximation based technique for doing parameter estimation under Ensemble Kalman filter set-up. Our proposed method is

dynamic in nature. At each time point, our method updates the parameter estimate using the last estimate and the observed data for the current time. Under some suitable conditions, we have showed that the estimated parameters converge to true values.

Using simulation studies we have showed it's performance in estimating model parameters, as well as it's relative strength compared to augmentation based parameter estimation technique. We have seen that, SA based approach works remarkably well for parameter estimation. Also in terms state estimation, our proposed approach results in much stable state estimation, relative to the augmentation based approach. We have noticed for more complicated model, like the Spatial RCA model, SA based approach fails to perform well in very high dimension. Finally we have applied our method to large spatio-temporal data.

3. NONPARAMETRIC SEEMINGLY UNRELATED REGRESSION WITH GAUSSIAN GRAPHICAL MODEL

3.1 Introduction

Seemingly Unrelated Regression(SUR), introduced by Zellner (1962), considers a set of seemingly unrelated regression equations with correlated errors. This kind of model is in general applicable to various fields, such as genetics, econometrics, sociology etc. Non-parametric SUR regression (see Smith and Kohn (2000), Holmes et al. (2002)) is a method of estimating SUR regression function without assuming any specific form for the mean function. In this paper we propose a Bayesian technique for consistently estimating the regression function in the case of relatively small sample size compared to the number of parameters to be estimated. The regression model considered in our paper is given by

$$\mathbf{y}_i = f_i(\mathbf{X}) + \boldsymbol{\epsilon}_{i\cdot}, \quad \forall \quad i = 1, \ldots, q,$$

where the subscript i denotes the i^{th} regression equation, $\mathbf{y}_i$, n×1, is the dependent variable, $\mathbf{X}$, n×p, is the predictor and f_i is the unknown regression function. Like other SUR model, we assume correlated Gaussian errors,

$$\boldsymbol{\epsilon}_{\cdot j} \overset{iid}{\sim} \mathrm{N}(0, \boldsymbol{\Sigma}), \quad \forall \quad j = 1, \ldots, n,$$

where $\boldsymbol{\epsilon}_{\cdot j}$, q×1, is the error vector corresponding to the j^{th} observation. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, be the corresponding precision matrix. Define, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_{1\cdot}, \cdots, \boldsymbol{\epsilon}_{q\cdot}) = (\boldsymbol{\epsilon}'_{\cdot 1}, \cdots, \boldsymbol{\epsilon}'_{\cdot n})'$. Then using the definition of Matrix Variate Normal(MVN)(see Dawid (1981)), we write

$$\boldsymbol{\epsilon} \sim \mathrm{MVN}(0, \mathbf{I}_n, \boldsymbol{\Sigma}). \tag{3.1}$$

We can restate our model as

$$\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\epsilon}, \tag{3.2}$$

where $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_q)$, n×q, is the dependent variable matrix and $f(\mathbf{X}) = (f_1(\mathbf{X}), \cdots, f_q(\mathbf{X}))$, n×q.

Various nonparametric techniques including smoothing spline, kernel smoothers (see Hastie and Tibshirani (1990)), spline with adaptive knots(see Friedman (1991)), variable bandwidth kernel method (see Fan and Gijbels (1995)) exists in the literature. Denison et al. (1998) introduced an adaptive piecewise polynomial function to estimate f_i's. In our paper, we use a *piecewise linear* function with fixed number of equally spaced knots to reduce the computational complexity. In our paper we use variable selection techniques to find a subset of unknown regressors (Holmes et al. (2002)).

In both Smith and Kohn (2000) and Holmes et al. (2002), $\boldsymbol{\Sigma}$ is treated as a nuisance parameter, and not much attention has been given in examining graphical structure of $\mathbf{Y}$. From portfolio management problem in finance, to enormous marketing data bases, to gene expression studies in bioinformatics, we are now faced with high dimensional problems with significant graphical structure. Gaussian graphical model has a long history of being used as a tool for estimating sparse Graphical structure under high dimensional set-up - see Carvalho and Scott (2009) and the references there. Rather than directly estimating the covariance matrix $\boldsymbol{\Sigma}$, we model $\boldsymbol{\Omega}$, the precision matrix. A good thing about the precision matrix, $\boldsymbol{\Omega}$, and graphical conditional independence. By modelling conditional independence, Gaussian graphical model makes the computation efficient and scalable by introducing sparsity in graph. Although both nonparametric SUR and Gaussian Graphical modelling have long history, joint estimation have never been done. Bhadra and Mallick (2013) introduces an efficient Bayesian algorithm for linear SUR. However, a number of criticisms can be directed at a linear Gaussian modeling approach. Perhaps the most severe among them is the fact that if the true mean function is non-

linear, a linear model will completely fail to capture that. Indeed, in most applications linearity assumption is imposed to reduce the computational burden. However, there are many situations where more flexibility may be desired. The present article achieves this in a *multiple predictor, multiple responses* Bayesian regression setting via the use of a piecewise linear spline basis function.

The chapter is organized as follows. In Section 2, we describe Gaussian Graphical model and introduce the Bayesian hierarchical model. Two simulations studies are done in section 3. In section 4 and 5, we display the efficiency of our method by comparing with other methods using real data. Section 6 gives a brief discussion of our findings.

3.2 The Model

3.2.1 Bayesian Graphical Model

An undirected graph $\mathbf{G}$ can be represented by the pair (V, E), where V represents the set of *vertices* and $E = (i, j)$ represents the set of edges, for some $i, j \in V$. Such an undirected covariance graph $\mathbf{G}$ can be used to model the conditional independence structure among the q-dimensional response variables in a Gaussian graphical model (Dempster (1972)). Two nodes, i and j, are called *n*eighbors if $(i, j) \in E$. A graph is called *complete*, if all possible pair of nodes are *n*eighbors. $C \subset \mathbf{G}$, is called *c*omplete if it induces a complete subgraph. A *c*omplete subset that is maximal is called a *c*lique (see Lauritzen (1996) for details). If two *c*liques overlap in a set S, then S is called a *s*eparator of those cliques. An undirected, marked graph is said to be *d*ecomposable if it is *c*omplete, or if there exists a triple (A, B, C) such that $V = A \cup B \cup C$, $C = A \cap B$ and C is a complete subset of V.

Properties of Gaussian distribution implies that if $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_q) \sim \text{MVN}_{n\times q}(\boldsymbol{\eta}, \mathbf{I}_n, \boldsymbol{\Sigma})$, then $\mathbf{y}_i$ and $\mathbf{y}_j$ are conditionally independent if and only if $\boldsymbol{\Omega}_{ij} = 0$. Thus, if $\mathbf{G}$ is the adjacency graph corresponding to the inverse covariance matrix $\boldsymbol{\Omega}$, then presence of an off-diagonal edge between two nodes imply non-zero partial correlation (i.e., conditional

dependence) and the absence of an edge imply conditional independence.

The hyper-inverse Wishart is the general set of conjugate priors that was introduced by Dawid and Lauritzen (1993). Together with the concept of decomposable graph, this set of priors are computationally efficient as shown in Jones et al. (2005), Carvalho and Scott (2009) and Bhadra and Mallick (2013). Suppose the decomposable graph, $\mathbf{G}$, can be split into a set of cliques, $\{c_1, \ldots, c_k\}$. Define the set of separator $\{s_2, \ldots, s_k\}$, where $s_j = (c_1 \cup \ldots \cup c_{j-1}) \cap c_j$. Then using equation (5.44) in Lauritzen (1996), we can write

$$f(\mathbf{y}) = \frac{\prod_{j=1}^{k} f(\mathbf{y}_{c_j})}{\prod_{j=2}^{k} f(\mathbf{y}_{s_j})}. \tag{3.3}$$

Also from Dawid and Lauritzen (1993) we know that, given $\mathbf{G}$, if $\mathbf{y}|\mathbf{\Sigma}_G \sim \text{MVN}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma}_G)$ and $\mathbf{\Sigma}_G|\mathbf{G} \sim \text{HIW}(\delta, \mathbf{\Phi})$, for some positive integer δ and positive definite matrix Φ,we have $\mathbf{\Sigma}_G|\mathbf{y}, \mathbf{G} \sim \text{HIW}(\delta + n, \mathbf{\Phi} + \mathbf{y}'\mathbf{y})$.

3.2.2 Hierarchical Model

From equations (3.1)-(3.2), we have

$$\mathbf{Y}|\mathbf{\Sigma}_G \sim \text{MVN}_{n\times q}(f(\mathbf{X}), \mathbf{I}_n, \mathbf{\Sigma}),$$

where the function $f : \text{R}^{n\times p} \to \text{R}^{n\times q}$ performs a smooth, nonlinear mapping from the p-dimensional predictor space to the q-dimensional response space. In our paper we

estimate the non-linear unknown function f by a piecewise linear spline. Define,

$$\mathbf{U}_{n\times p(k+1)} = \begin{bmatrix} X_{1.} & (X_{1.}-w_1)_+ & \dots & (X_{1.}-w_k)_+ \\ X_{2.} & (X_{2.}-w_1)_+ & \dots & (X_{2.}-w_k)_+ \\ \vdots & \vdots & \vdots & \ddots \\ X_{n.} & (X_{n.}-w_1)_+ & \dots & (X_{n.}-w_k)_+ \end{bmatrix}$$

$$\mathbf{B}_{p(k+1)\times q} = \begin{bmatrix} \beta_{110} & \beta_{210} & \dots & \beta_{q10} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{1p0} & \beta_{2p0} & \dots & \beta_{qp0} \\ \beta_{111} & \beta_{211} & \dots & \beta_{q11} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{1p1} & \beta_{2p1} & \dots & \beta_{qp1} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{1pk} & \beta_{2pk} & \dots & \beta_{qpk} \end{bmatrix}$$

where $\mathbf{X}_{i.} = (x_{i1}, \dots, x_{ip}), \forall\ i = 1, \dots, n$, $w_1, \dots, w_k$ are k equally spaced knot points, $(\mathbf{X}_{i.} - w_l)_+ = [(x_{i1} - w_l)_+, \dots, (x_{ip} - w_l)_+], \forall\ i, l$. Here p is the total number of covariates. We approximate f by

$$f(\mathbf{X}) = \mathbf{UB}.$$

To introduce the notion of redundant variables in $\mathbf{X}$, we define a binary vector $\gamma' = (\gamma_1, \dots, \gamma_p)$, where $\gamma_i = 0$ if $\beta_{ji0} = 0$ and $\beta_{jil} = 0, \forall j = 1, \dots, q, \quad l = 1, \dots, k$. Similarly, to introduce the notion of sparsity in the precision matrix, we define the binary variable $G_l, \quad l = 1, \dots, \frac{q(q-1)}{2}$, the l^{th} off diagonal element in the adjacency matrix corresponding to the graph $\mathbf{G}$. Diagonal elements of the adjacency matrix are restricted to 1.The bayesian hierarchical model is given

by

$$
\begin{aligned}
(\mathbf{Y} - \mathbf{U}_\gamma B_\gamma)|\boldsymbol{\gamma}, \boldsymbol{\Sigma}_G &\sim \text{MN}(0, \mathbf{I}_n, \boldsymbol{\Sigma}_G), && (3.4)\\
\mathbf{B}_\gamma|\boldsymbol{\gamma}, \boldsymbol{\Sigma}_G &\sim \text{MN}(0, c\mathbf{I}_{p_\gamma(k+1)}, \boldsymbol{\Sigma}_G) && (3.5)\\
\boldsymbol{\Sigma}_G|\mathbf{G} &\sim \text{HIW}_G(b, d\mathbf{I}_q), && (3.6)\\
\gamma_i &\overset{iid}{\sim} \text{Bernoulli}(w_\gamma) \text{ for } i = 1, \ldots, p, && (3.7)\\
G_l &\overset{iid}{\sim} \text{Bernoulli}(w_G) \text{ for } l = 1, \ldots, \frac{q(q-1)}{2}, && (3.8)\\
w_\gamma, w_G &\sim \text{U}(0, 1) && (3.9)
\end{aligned}
$$

where $\mathbf{U}_\gamma$ is the design matrix only containg regressors present in $\boldsymbol{\gamma}$, b, c, d are fixed hyper parameters and w's are used to control the sparsity of $\boldsymbol{\gamma}$ and $\mathbf{G}$. Also, denote $\mathbf{p}_\gamma = \sum \gamma_i$. From equation (3.5) we have

$$\mathbf{U}_\gamma \mathbf{B}_\gamma|\boldsymbol{\gamma}, \boldsymbol{\Sigma}_G \sim \text{MN}(0, c\mathbf{U}_\gamma \mathbf{U}'_\gamma, \boldsymbol{\Sigma}_G). \tag{3.10}$$

Hence from equation (3.4) we have

$$\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\Sigma}_G \sim \text{MN}(0, \mathbf{I}_n + c\mathbf{U}_\gamma \mathbf{U}'_\gamma, \boldsymbol{\Sigma}_G). \tag{3.11}$$

Let $\mathbf{L}\mathbf{L}' = (\mathbf{I}_n + c\mathbf{U}_\gamma \mathbf{U}'_\gamma)^{-1}$, where $\mathbf{L}$ is the corresponding Cholesky decomposition. Define $\mathbf{T} = \mathbf{L}\mathbf{Y}$. Then

$$\mathbf{T}|\boldsymbol{\gamma}, \boldsymbol{\Sigma}_G \sim \text{MN}(0, \mathbf{I}_n, \boldsymbol{\Sigma}_G). \tag{3.12}$$

Integrating $\boldsymbol{\Sigma}_G$ from equation (3.12), and using equation (3.8) we get

$$\mathbf{T}|\boldsymbol{\gamma}, \mathbf{G} \sim \text{HMT}_{n\times q}(b, \mathbf{I}_n, d\mathbf{I}_q), \tag{3.13}$$

where HMT() is the Hyper Matrix t defined in Dawid and Lauritzen (1993). Given the graph, distribution of Hyper Matrix t can be looked upon as products and ratios over the cliques and separators similar to equation (3.3). Suppose we have n observations and the graph $\mathbf{G}$ is given

with the set of cliques $\{c_1, \ldots, c_k\}$ and separators $\{s_2, \ldots, s_k\}$. Given $A \in c_j$, consider the nodes in A, and denote $\mathbf{T}_A$ as the corresponding $n \times |A|$ matrix. Density of Hyper Matrix t on the given clique c_j, is given by

$$f(\mathbf{t}_{c_j}) = \pi^{-n/2} \frac{\Gamma_{|c_j|}((b+n+|c_j|-1)/2)}{\Gamma_{|c_j|}((b+|c_j|-1)/2)} |dI_{|c_j|}|^{-n/2} |I_n + (\mathbf{t}_{c_j})(dI_{|c_j|})^{-1}(\mathbf{t}_{c_j})'|^{-(b+n+|c_j|-1)/2}. \tag{3.14}$$

By combining equations (3.3) and (3.14) we can derive the complete density of $\mathbf{t}$. For the detailed description of the Hyper Matrix t density function, see equation (46) in Dawid and Lauritzen (1993).

3.2.3 Null-Based Bayes Factor

In the null-based approach of calculating Bayes factor we compare null model $\mathcal{M}_0$ against the alternative model $\mathcal{M}_A$. Let $\mathcal{M}_0 = (\mathbf{\Gamma}_0, \mathbf{G}_0)$ be the null model where $\boldsymbol{\gamma}_0$ indicating no regressor being selected and $\mathbf{G}_0$ denotes the null graph with no edges and $\mathcal{M}_A = (\mathbf{\Gamma}_A, \mathbf{G}_A)$ is the alternative model. We assume a priori $\Pr(\mathcal{M}_0) = \Pr(\mathcal{M}_A) = 1/2$. The null-based Bayes factor, given by

$$\mathrm{BF}(\mathcal{M}_0, \mathcal{M}_A) = \frac{f(\mathbf{t}|\mathcal{M}_0)}{f(\mathbf{t}|\mathcal{M}_A)} = \frac{f(\mathbf{t}|\mathbf{\Gamma}_0, \mathbf{G}_0)}{f(\mathbf{t}|\mathbf{\Gamma}_A, \mathbf{G}_A)}, \tag{3.15}$$

where $\mathbf{t} = (\mathbf{I}_n + c(\mathbf{X}_\gamma \mathbf{X}_\gamma'))^{-\frac{1}{2}}\mathbf{y}$, is used to test the above hypothesis. Now Equation (3.15) can be further decomposed as

$$\mathrm{BF}(\mathcal{M}_0, \mathcal{M}_A) = \mathrm{BF}(\mathbf{\Gamma}_0, \mathbf{G}_0; \mathbf{\Gamma}_A, \mathbf{G}_0) \times \mathrm{BF}(\mathbf{\Gamma}_A, \mathbf{G}_0; \mathbf{\Gamma}_A, \mathbf{G}_A). \tag{3.16}$$

Consider nested sequence $\mathbf{\Gamma}_0 \subset \mathbf{\Gamma}_1 \subset \cdots \subset \mathbf{\Gamma}_A$ of models that differs by a single regressor. With out loss of generality, assume that $\mathbf{\Gamma}_i = (\gamma_1, \ldots, \gamma_p)$, where $\gamma_j = 1$, $j = 1, \ldots, i$ and $\gamma_j = 0$, $j = i+1, \ldots, p$. Also consider the sequence $\mathbf{G}_0 \subset \mathbf{G}_1 \subset \cdots \subset \mathbf{G}_d = \mathbf{G}_A$ of decomposable graphs that differ only by one edge. Let e_i denote the edge in $\mathbf{G}_i$ but not in $\mathbf{G}_{i-1}$, and let C_i be the unique clique of $\mathbf{G}_i$ containing e_i.

$$
\begin{aligned}
\mathrm{BF}(\mathbf{\Gamma}_0, \mathbf{G}_0; \mathbf{\Gamma}_A, \mathbf{G}_0) &= \frac{F(\mathbf{t}|\mathbf{\Gamma}_0, \mathbf{G}_0)}{f(\mathbf{t}|\mathbf{\Gamma}_A, \mathbf{G}_0)} = \prod_{i=0}^{A-1} \frac{f(\mathbf{t}|\mathbf{\Gamma}_i, \mathbf{G}_0)}{f(\mathbf{t}|\mathbf{\Gamma}_{i+1}, \mathbf{G}_0)} \\
&= \prod_{i=0}^{A-1} \prod_{j=1}^{q} \frac{f(\mathbf{t}_{i,j})}{f(\mathbf{t}_{i+1,j})}, \\
&\quad \text{where } \mathbf{t}_{i,j} = (\mathbf{I}_n + c\mathbf{X}_{\Gamma_i}\mathbf{X}'_{\Gamma_i})^{-\frac{1}{2}}\mathbf{y}_j, \mathbf{y}_j \text{ is the j}^{\text{th}} \text{ column of } \mathbf{Y}. \\
&= \prod_{i=0}^{A-1} \prod_{j=1}^{q} \frac{\pi^{-n/2}\frac{\Gamma(\frac{b+n}{2})}{\Gamma(\frac{b}{2})}[\det(\mathbf{I}_n + \mathbf{t}_{i,j}\mathbf{t}'_{i,j}/d)]^{-(b+n)/2}}{\pi^{-n/2}\frac{\Gamma(\frac{b+n}{2})}{\Gamma(\frac{b}{2})}[\det(\mathbf{I}_n + \mathbf{t}_{i+1,j}\mathbf{t}'_{i+1,j}/d)]^{-(b+n)/2}} \\
&\quad \text{using equations (3.3) and (3.14)} \\
&= \prod_{i=0}^{A-1} \prod_{j=1}^{q} [\frac{\det(\mathbf{I}_n + \mathbf{t}_{i,j}\mathbf{t}'_{i,j}/d)}{\det(\mathbf{I}_n + \mathbf{t}_{i+1,j}\mathbf{t}'_{i+1,j}/d)}]^{-(b+n)/2} \\
&= \prod_{i=0}^{A-1} \prod_{j=1}^{q} \tau_{i,j}^{-(b+n)/2} = \prod_{i=0}^{A-1} \tau_i^{-(b+n)/2}, \qquad (3.17)
\end{aligned}
$$

where $\tau_{i,j} = \det(\mathbf{I}_n + \mathbf{t}_{i,j}\mathbf{t}'_{i,j}/d)/\det(\mathbf{I}_n + \mathbf{t}_{i+1,j}\mathbf{t}'_{i+1,j}/d)$. Here $\mathbf{X}_{\Gamma_{i+1}} = (\mathbf{X}'_{\Gamma_i}, \mathbf{x}'_{i+1})'$, where n×1 predictor $\mathbf{x}_{i+1}$ is present in $\mathbf{X}_{\Gamma_{i+1}}$ bur not in $\mathbf{X}_{\Gamma_i}$.

Consider 2 graphs, $\mathbf{G}_0^*$ and $\mathbf{G}^*$, where $\mathbf{G}_0^*$ has exactly one edge $e = \{\gamma, \mu\}$ less than $\mathbf{G}^*$. Let $\mathbf{C}^* = (\mu, 1, \ldots, k, \eta)$ be the unique clique of of $\mathbf{G}$ containing e. Suppose $\mathbf{C}_0^*$, $\mathbf{C}_1^*$ and $\mathbf{S}^*$ are such that $\mathbf{C}_0^* = \mathbf{C}^* \backslash \{\eta\}$, $\mathbf{C}_1^* = \mathbf{C}^* \backslash \{\mu\}$ and $\mathbf{S}^* = \mathbf{C}^* \backslash \{\eta, \mu\}$. The Bayes factor, comparing

$(\mathbf{\Gamma}_A, \mathbf{G}_0^*)$ to $(\mathbf{\Gamma}_A, \mathbf{G}^*)$ is given by

$$\begin{aligned}
\mathrm{BF}(\mathbf{\Gamma}_A, \mathbf{G}_0^*; \mathbf{\Gamma}_A, \mathbf{G}^*) &= \frac{f(\mathbf{t}|\mathbf{\Gamma}_A, \mathbf{G}_0^*)}{f(\mathbf{t}|\mathbf{\Gamma}_A, \mathbf{G}^*)} \\
&= \mathrm{K}_0 \cdot \frac{|\mathbf{I}_n + \frac{t_{\mathbf{S}^*} t'_{\mathbf{S}^*}}{d}|^{\frac{b+n+k-1}{2}} |\mathbf{I}_n + \frac{t_{\mathbf{C}^*} t'_{\mathbf{C}^*}}{d}|^{\frac{b+n+k+1}{2}}}{|\mathbf{I}_n + \frac{\mathbf{t}_{\mathbf{C}_0^*} \mathbf{t}'_{\mathbf{C}_0^*}}{d}|^{\frac{b+n+k}{2}} |\mathbf{I}_n + \frac{\mathbf{t}_{\mathbf{C}_1^*} \mathbf{t}'_{\mathbf{C}_1^*}}{d}|^{\frac{b+n+k}{2}}} \\
&\quad \text{using equations (3.3) and (3.14)} \\
&= \mathrm{K}_0 \cdot \frac{|\mathbf{I}_k + \frac{\mathbf{t}'_{\mathbf{S}^*} \mathbf{t}_{\mathbf{S}^*}}{d}|^{\frac{b+n+k-1}{2}} |\mathbf{I}_{k+2} + \frac{\mathbf{t}'_{\mathbf{C}^*} \mathbf{t}_{\mathbf{C}^*}}{d}|^{\frac{b+n+k+1}{2}}}{|\mathbf{I}_{k+1} + \frac{\mathbf{t}'_{\mathbf{C}_0^*} \mathbf{t}_{\mathbf{C}_0^*}}{d}|^{\frac{b+n+k}{2}} |\mathbf{I}_{k+1} + \frac{\mathbf{t}'_{\mathbf{C}_1^*} \mathbf{t}_{\mathbf{C}_1^*}}{d}|^{\frac{b+n+k}{2}}} \\
&= \mathrm{K}_0 \cdot \left[\frac{|\mathbf{I}_k + \frac{\mathbf{t}'_{\mathbf{S}^*} \mathbf{t}_{\mathbf{S}^*}}{d}| \cdot |\mathbf{I}_{k+2} + \frac{\mathbf{t}'_{\mathbf{C}^*} \mathbf{t}_{\mathbf{C}^*}}{d}|}{|\mathbf{I}_{k+1} + \frac{\mathbf{t}'_{\mathbf{C}_0^*} \mathbf{t}_{\mathbf{C}_0^*}}{d}| \cdot |\mathbf{I}_{k+1} + \frac{\mathbf{t}'_{\mathbf{C}_1^*} \mathbf{t}_{\mathbf{C}_1^*}}{d}|} \right]^{\frac{b+n+k+1}{2}} \\
&\qquad \times \frac{|\mathbf{I}_{k+1} + \frac{\mathbf{t}'_{\mathbf{C}_0^*} \mathbf{t}_{\mathbf{C}_0^*}}{d}| \cdot |\mathbf{I}_{k+1} + \frac{\mathbf{t}'_{\mathbf{C}_1^*} \mathbf{t}_{\mathbf{C}_1^*}}{d}|}{|\mathbf{I}_k + \frac{\mathbf{t}'_{\mathbf{S}^*} \mathbf{t}_{\mathbf{S}^*}}{d}|^2}, \qquad (3.18)
\end{aligned}$$

where K_0 is the appropriate constant. So the Bayes factor comparing $(\mathbf{\Gamma}_A, \mathbf{G}_0)$ to $(\mathbf{\Gamma}_A, \mathbf{G}_a)$ can be calculated as

$$\begin{aligned}
\mathrm{BF}(\mathbf{\Gamma}_A, \mathbf{G}_0; \mathbf{\Gamma}_A, \mathbf{G}_A) &= \frac{f(\mathbf{t}|\mathbf{\Gamma}_A, \mathbf{G}_0)}{f(\mathbf{t}|\mathbf{\Gamma}_A, \mathbf{G}_A)} \\
&= \prod_{i=0}^{d-1} \frac{f(\mathbf{t}|\mathbf{\Gamma}_A, \mathbf{G}_i)}{f(\mathbf{t}|\mathbf{\Gamma}_A, \mathbf{G}_{i+1})} = \prod_{i=0}^{d-1} \mathrm{BF}(\mathbf{\Gamma}_A, \mathbf{G}_i; \mathbf{\Gamma}_A, \mathbf{G}_{i+1}). \qquad (3.19)
\end{aligned}$$

Using (3.16), (3.17) and (3.19), we can get the null based Bayes factor.

3.2.4 MCMC for γ given G and T

From equation (3.7), we have $p(\gamma|w_\gamma) = w_\gamma^{p_\gamma}(1-w_\gamma)^{p-p_\gamma}$, $p_\gamma = 1, \ldots, p$ where w_γ is U(0,1) and p_γ is the number of non-zero elements of $\boldsymbol{\gamma}$. Integrating w_γ out, we get $p(\boldsymbol{\gamma}) = \mathrm{Beta}(\mathbf{p}_\gamma, \mathbf{p} - \mathbf{p}_\gamma + 1)$. Hence the sampling procedure proceeds as follows:

1. Given $\boldsymbol{\gamma}$, propose $\boldsymbol{\gamma}^*$ by either changing a non-zero entry to zero with probability (1-u) and set $q(\gamma|\gamma^*)/q(\gamma^*|\gamma) = \frac{u}{1-u}$, or changing a zero entry in γ to one, with probability u and set $q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)/q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}) = \frac{1-u}{u}$.

2. Calculate $f(\mathbf{t}|\boldsymbol{\gamma}^*, \mathbf{G})$ and $f(\mathbf{t}|\boldsymbol{\gamma}, \mathbf{G})$ where f denotes the HMT density, derived by com-

bining equations (3.3) and (3.14).

3. Accept $\boldsymbol{\gamma}^*$ with probability

$$r(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) = \min\big(1, \frac{f(\mathbf{t}|\boldsymbol{\gamma}^*, \mathbf{G})p(\boldsymbol{\gamma}^*)q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{f(\mathbf{t}|\boldsymbol{\gamma}, \mathbf{G})p(\boldsymbol{\gamma})q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})}\big)$$

3.2.5 MCMC for G given $\boldsymbol{\gamma}$ and T

From equation 3.8, we have $p(\mathbf{G}|w_G) = w_G^g(1-w_G)^{q(q-1)/2-g}$, where $g = 1, \ldots, q(q-1)/2$. Since w_G is U(0,1), integrating out w_G we get $p(\mathbf{G}) = \text{Beta}(g+1, \frac{q(q-1)}{2} - g + 1)^{-1}$, where g is the number of non-zero off diagonal elements in $\mathbf{G}$. Here the MCMC works as follows

1. Given $\mathbf{G}$, propose decomposable graph $\mathbf{G}^*$ by either changing a non-zero off-diagonal entry in the lower triangular part of $\mathbf{G}$ to zero with probability (1-u) and set $q(\mathbf{G}|\mathbf{G}^*)/q(\mathbf{G}^*|\mathbf{G}) = \frac{u}{1-u}$, or changing a zero off-diagonal entry in the lower triangular part of $\mathbf{G}$ to one, with probability u and set $q(\mathbf{G}|\mathbf{G}^*)/q(\mathbf{G}^*|\mathbf{G}) = \frac{1-u}{u}$. We change the corresponding upper triangular matrix in a consistent fashion.

2. Calculate $f(\mathbf{t}|\boldsymbol{\gamma}, \mathbf{G}^*)$ and $f(\mathbf{t}|\boldsymbol{\gamma}; \boldsymbol{\lambda}; G)$ where f denotes the HMT density, derived by combining equations (3.3) and (3.14).

3. Accept $\mathbf{G}^*$ with probability

$$r(\mathbf{G}, \mathbf{G}^*) = \min\big(1, \frac{f(\mathbf{t}|\boldsymbol{\gamma}, \mathbf{G}^*)p(\mathbf{G}^*)q(\mathbf{G}|\mathbf{G}^*)}{f(\mathbf{t}|\boldsymbol{\gamma}, \mathbf{G})p(\boldsymbol{\gamma})q(\mathbf{G}^*|\mathbf{G})}\big)$$

3.2.6 Sampling B_γ and Σ_G from its Posterior

The coefficient matrix $\mathbf{B}_\gamma$ can not be fully recovered. We can only recover the globally significant ones conditional on $\mathbf{Y}$, $\boldsymbol{\gamma}$, and $\mathbf{G}$ as follows:

1. Generate $\boldsymbol{\Sigma}_G|\mathbf{Y}, \mathbf{B}_\gamma, \boldsymbol{\gamma}, \mathbf{G}$ from $\text{HIW}(b+n, d\mathbf{I}_q + (\mathbf{Y} - \mathbf{U}_\gamma\mathbf{B}_\gamma)'(\mathbf{Y} - \mathbf{U}_\gamma\mathbf{B}_\gamma))$.

2. Generate $\mathbf{B}_\gamma|\mathbf{Y}, \boldsymbol{\Sigma}_G, \boldsymbol{\gamma}, \mathbf{G}$from $\text{MN}((\mathbf{U}_\gamma'\mathbf{U}_\gamma + c^{-1}\mathbf{I}_{p_\gamma k})^{-1}\mathbf{U}_\gamma'\mathbf{Y}, (\mathbf{U}_\gamma'\mathbf{U}_\gamma + c^{-1}\mathbf{I}_{p_\gamma k})^{-1}, \boldsymbol{\Sigma}_G)$.

Since generating $\mathbf{B}_\gamma$ can be a problem because of the dimensionality, we define,

$$\mathbf{H}_\gamma = \mathbf{B}_\gamma \boldsymbol{\Sigma}_G^{-\frac{1}{2}},$$

where $\boldsymbol{\Sigma}_G^{-\frac{1}{2}}$ is a Cholesky Decomposition of $\boldsymbol{\Sigma}_G^{-1}$. Then

1. Generate $\boldsymbol{\Sigma}_G|\mathbf{Y}, \mathbf{B}_\gamma, \boldsymbol{\gamma}, \mathbf{G}$ as specified above.
2. Generate $\mathbf{H}_\gamma|\mathbf{Y}, \boldsymbol{\Sigma}_G, \boldsymbol{\gamma}, \mathbf{G}$ from $\text{MN}((\mathbf{U}_\gamma'\mathbf{U}_\gamma + c^{-1}\mathbf{I}_{p_\gamma k})^{-1}\mathbf{U}_\gamma'\mathbf{Y}\boldsymbol{\Sigma}_G^{-\frac{1}{2}}, (\mathbf{U}_\gamma'\mathbf{U}_\gamma + c^{-1}\mathbf{I}_{p_\gamma k})^{-1}, \mathbf{I}_q)$.
3. $\mathbf{B}_\gamma|\mathbf{Y}, \boldsymbol{\Sigma}_G, \boldsymbol{\gamma}, \mathbf{G} = \mathbf{H}_\gamma \boldsymbol{\Sigma}_G^{\frac{1}{2}}$.

3.2.7 Choosing the Hyper-parameters

One of the obvious question one might ask is how to choose the hyper-parameters b, c and d. In general choice of b is not that important but both choice of c and d plays very important role. d works as a shrinkage parameter for the graph where as c plays a global shrinkage parameter similar to ridge regression. To reduce the role of d in our analysis and make the analysis free of tuning the values of d, we did the following (see Liu and Wang (2012) for more details)

- rather than using $\mathbf{Y}$ for our data analysis, we used $\tilde{\mathbf{Y}} = \mathbf{Y}\hat{\Sigma}_{\mathbf{Y}}^{-1/2}$;
- then we took $d = \sqrt{\frac{\ln(q)}{2n}}$.

The other hyper-parameter c is chosen in a way to match the variability in the data, ie., $\tilde{\mathbf{Y}}$ and the design matrix $\mathbf{U}$. We set $c = \lambda * \frac{\text{Var}(\text{vec}(\tilde{\mathbf{Y}}))}{\text{Var}(\text{vec}(\mathbf{U}))}$, where λ can be chosen from [0.1,10].

3.3 Simulation Study

Here we present two simulation studies. In the first simulation, we consider a small dimensional problem, and show that the variable selection and graph selection works quite well. Also, for this problem, we have regenerated $\mathbf{B}$ and $\boldsymbol{\Sigma}_G$ to show that the predicted values can mimic the non-linear shape of the observations. For the second simulation, we considered a high dimensional problem, and showed that our method works well to identify the regressor variables as well as the underlying graph.

3.3.1 Simulation One

First we start with a relatively small dimensional problem, with $p = 10$, $q = 50$ and $n = 120$. $\mathbf{x}$ is simulated randomly from $\mathrm{U}(-2, 2)$. We used a relatively simple, smooth function, similar to the function used in Denison et al. (1998),

$$f_1(\mathbf{x}) = \sin(2x_2) + 2e^{-16x_6^2} \tag{3.20}$$

to generate data. The function is re-scaled such that its support is the unit interval. The dependent variable, $\mathbf{Y}$, is generated by adding zero mean Gaussian error to re-scaled $f_1(x)$. The error adjacency matrix, corresponding to the true graph $\mathbf{G}$, is shown in Figure 3.1.

For this simulation, we have used fifteen fixed knot-points. 20,000 MCMC iteration steps are performed, after 10,000 burn-in steps. Figure 3.1 shows the estimated adjacency matrix and the estimated posterior probability of $\mathbf{p}_\gamma$. Using 0.5 as our cut-off on the estimated posterior probability, we have selected x_2 and x_6. By using the linear model described in Bhadra and Mallick (2013), we get the estimated adjacency matrix and the estimated posterior probability of $\mathbf{p}_\gamma$, showed in Figure 3.2. Although the linear model identifies the correct graph, it clearly fails to separate the true regressor variables from the false ones.

Next we generate $\mathbf{B}_\gamma$ and $\mathbf{\Sigma}_G$, given identified regressors and the estimated adjacency matrix. The scatter plot in Figure 3.3 shows that our model can recapture the non-linear shape. The root mean square error, averaged over 10 independent MCMC runs, is 0.3363. The RMSE for the linear model, averaged over 10 independent MCMC runs, is 0.3379.

3.3.2 Simulation Two

For the second simulation set-up, we consider a high dimensional case, where $p = 100$, $q = 100$ and $n = 120$. X, 120×100, is simulated randomly from $\mathrm{U}(-2, 2)$. We used the following function,

$$f_2(\mathbf{x}) \quad = \quad \beta_1 \sin(x_{30}) + \beta_2 \sin(x_{50}) + \beta_3 x_{56} + \beta_4 \exp(\frac{x_{75}}{2}), \tag{3.21}$$

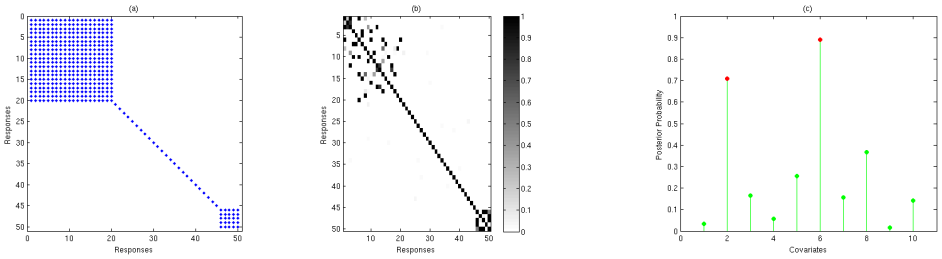

Figure 3.1: (a) True Adjacency Matrix. (b) Estimated Adjacency Matrix. (c)Posterior Probability Plot for $\boldsymbol{\gamma}$. Variable marked by red circle are the true variables identified by our model.

to generate data. β_i's, $i = 1, \ldots, 4$, are generated randomly from a zero mean multivariate Gaussian distribution. The observation, $\mathbf{Y}$, 120×100, is generated by adding adding zero mean Gaussian error to $f_2(x)$. The error adjacency matrix is shown in Figure 3.4(a). Estimated adjacency matrix and the variables are shown in Figure 3.4(b)-(c).

3.4 Scottish Elections

Scottish election data has earlier been used in various multivariate regression studies (Brown (1980),Breiman and Friedman (1997),Holmes et al. (2002)). The data consists of electoral results for all 71 Scottish constituencies in the United Kingdom general elections of February and October 1974. The raw data consists of the total votes for each of the four major parties (Conservative, Labour, Liberal and Nationalist) in each election, together with a categorical variable

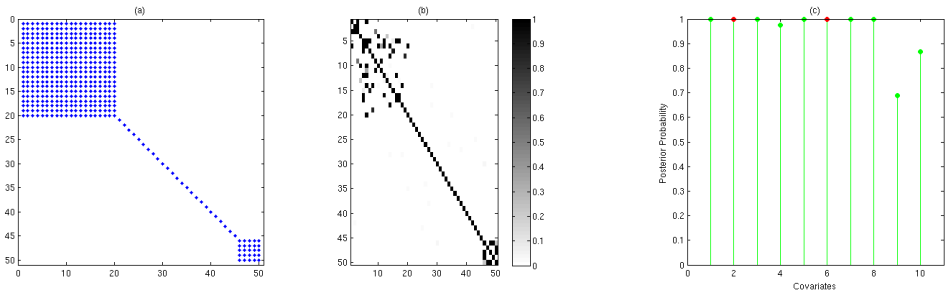

Figure 3.2: (a) True Adjacency Matrix. (b) Estimated Adjacency Matrix. (c)Posterior Probability Plot for $\boldsymbol{\gamma}$. Variable marked by red circle are the true variables.

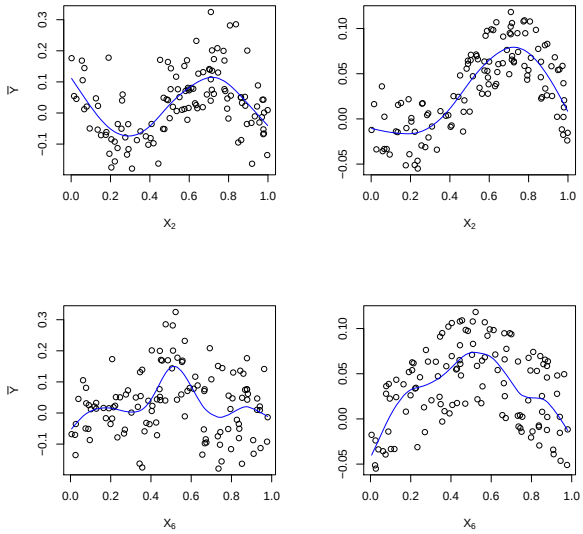

Figure 3.3: Left panel shows he scatter plot of Y, averaged over it's dimension, plotted against x_2 and x_6. Panel on the right shows the predicted values of Y, averaged over dimension, plotted against x_2 and x_6. Blue lines are the LOESS lines.

listing the location of the constituency by six regions, and the size of the electorate in each constituency. From the raw data the four response variable and the seven independent/regressor variable are derived as

(a) y=(y_1,y_2,y_3,y_4) is the difference between the vote share of four parties (Conservative, Labour, Liberal and Nationalist) in October and February.

(b) x_1, x_2, x_3 and x_4 are the vote share of the four parties in February.

(c) $x_5 = 0.5$ if Liberal intervenes (Liberal vote in February = 0; Liberal vote ¿ 0 in October); else $x_5 = 0$.

(d) $x_6 = 0.5$ if rural constituency (R=5,6); else $x_6 = 0$.

(e) $x_7 = 0.5$ if Labour or Nationalist won in February and $|x_2 - x_4| < 0.2$; else $x_7 = 0$.

First 30 data points are used for developing the model and the rest 41 data points are used as an out-of-sample test data set.

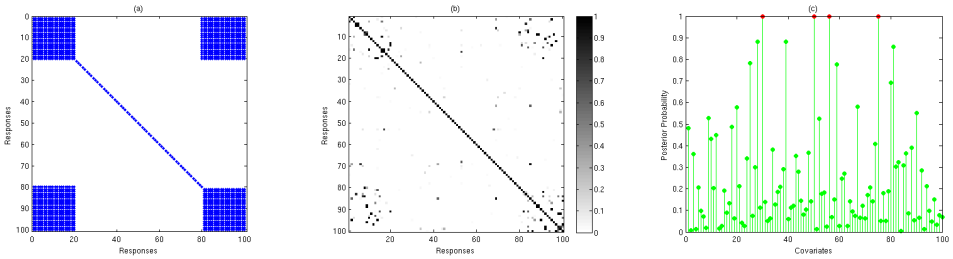

Figure 3.4: (a) True Adjacency Matrix. (b) Estimated Adjacency Matrix. (c)Posterior Probability Plot for γ. Variable marked by red circle are the true variables identified by the model.

We first calculated the sample partial correlation matrix and the corresponding p-values (it tests whether the corresponding partial correlation is significantly different from 0 or not) for the response variable. The p-values are given by

$$\begin{bmatrix} 0 & 0.7638 & 0.0029 & 0.0005 \\ & 0 & 0.0018 & 0.5916 \\ & & 0 & 0.0005 \\ & & & 0 \end{bmatrix}.$$

From above we can infer that (y_1, y_2) and (y_2, y_4) are partially independent. Similarly we calculated the correlation and the corresponding p-values between four response variable and the seven regressor variables. The p-values indicate that only x_3 and x_5 can be considered as the master predictors.

We model the data using our method with multivariate linear spline. We run the MCMC iteration 20000 times, with 10000 burn-in iterations.The posterior probability of the adjacency

matrix (same as graph $\mathbf{G}$) in this case is

$$\begin{pmatrix} 1.0000 & 0.3935 & 0.6683 & 1.0000 \\ & 1.0000 & 0.9224 & 0.2708 \\ & & 1.0000 & 0.6873 \\ & & & 1.0000 \end{pmatrix}.$$

The posterior probability plot of γ in Fig. 3.5 indicates that only x_3 and x_5 are the master predictor variables. Our MCMC result is consistent with the emperical findings above.

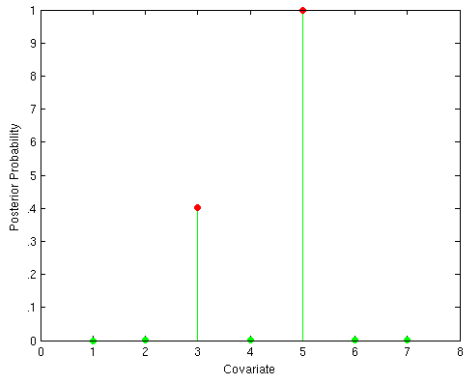

Figure 3.5: Estimated Posterior Probability of the Covariates.

The out-of-sample test results in terms of mean-squared prediction error multiplied by 1000, averaged over 5 MCMC run, is given in Table 3.1. Our overall model performance is comparable to all other models. Average MSE(*1000) in all our models are less than that in Breiman and Friedman (1997) and Holmes et al. (2002).

3.5 Asset Returns

Next we apply our method to analyze financial data. Here we have considered the weekly log return data set of 9 stocks from 2004, used earlier for analysis in Yuan et al. (2007) and Rothman et al. (2010). We are using this data to compare our method to the other methods.

Response	Result from various methods				
	(a)	(b)	(c)	(d)	(e)
1	0.98	0.98	0.82	0.83	0.80
2	0.58	0.44	0.35	0.35	0.35
3	0.38	1.26	0.50	0.60	0.43
4	1.92	1.17	2.06	2.04	1.94
Average	0.97	0.96	.93	0.95	0.88

Table 3.1: Predictive Mean Squared Error (times 1000) for Out-of-Sample Data in the Scottish Election Example. (a) corresponds to results in Breiman and Friedman (1997); (b) corresponds to Multivariate results in Holmes et al. (2002); (c)corresponds to a Linear model with only Graph selection; (d) corresponds to our full model with Non-parametric spline; (e) corresponds to a model with only Graph selection, but no variable selection,with Non-parametric spline.

We have used a vector AR(1) model, given by,

$$\mathbf{y} = f(\tilde{\mathbf{y}}) + \mathbf{E}, \tag{3.22}$$

where $\mathbf{y} = (y_2, \ldots, y_T)'$ and $\tilde{\mathbf{y}} = (y_1, \ldots, y_{T-1})'$. Here y_t is the vector of log-returns for week t. Following the earlier works, we have divided the data into training and validation data sets. The training data set is consist of first 26 weeks of data and the validation data set is the remaining 26 weeks of log-returns. We use the training data set to fit our model and then measure the Mean square error(MSE) for each stock for the validation data set.

We apply our model with multivariate linear spline on the training data. The model ran for 75,000 MCMC iterations together with 25,000 burn-ins. Fig. 3.6 shows the estimated adjacency matrix and the posterior probability of the covarites. 0.1 is used as the cut-off for covariate probabilities. From the figure, we conclude that the previous weeks log-return of Ford, Citi and AIG stock prices can be used as master predictor for our model. The cut-off used in case of adjacency matrix is 0.15. From our estimated adjacency matrix we find that companies with similar products are partially correlated - GM and Ford(car makers), Exxon and ConocoPhillips(oil and gas), Citi and AIG(financial), IBM and GE(technology). This result is consistent with subject knowledge.

Stock	Result from various methods			
	MRCE	ap.MRCE	FES	BGGM
Walmart	0.41	0.41	0.40	0.43
Exxon	0.31	0.31	0.29	0.32
GM	0.71	0.69	0.62	0.62
Ford	0.77	0.77	0.69	0.58
GE	0.45	0.45	0.41	0.40
ConocoPhillips	0.79	0.78	0.79	0.82
Citigroup	0.62	0.62	0.59	0.66
IBM	0.49	0.47	0.51	0.46
AIG	1.88	1.88	1.74	1.95
Average	0.71	0.71	0.67	0.69

Table 3.2: Mean Squared Error for each stock$\times$1000 based on the validation data. Results of MRCE and ap.MRCE methods are reported from Table 6 in Rothman et al. (2010) and that of FES method from Table 3 in Yuan et al. (2007). BGGM is our method.

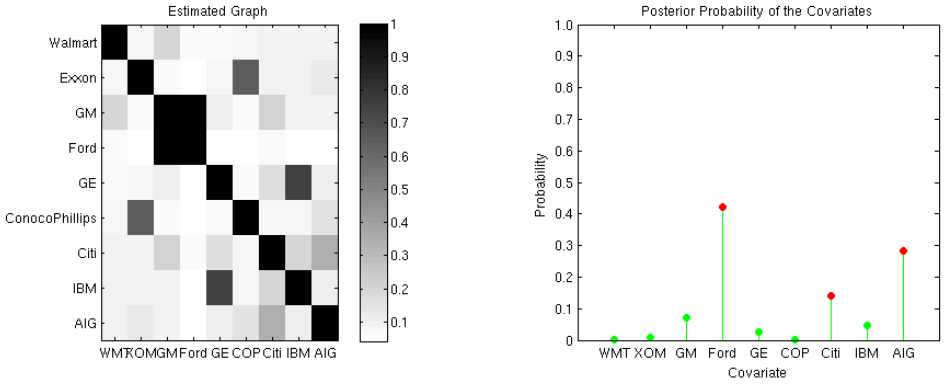

Figure 3.6: Estimated Graph and Posterior Probability of Covariates

Finally we use the estimated adjacency matrix and master covariates to estimate the spline coefficient matrix $\mathbf{B}$ based on the training data and use the estimated model to forecast for the validation data set. In table 3.2, we have reported average MSE for our model based on 10 MCMC runs, together with that from Yuan et al. (2007) and Rothman et al. (2010). We see that our result is comparable with both methods.

3.6 Discussion

The section discusses a joint variable and covariance selection technique for the case of a nonlinear, Gaussian graphical model. Relaxing the linearity assumption of Bhadra and Mallick (2013) allows one the important flexibility of being able to capture non-linear signals in the data. Ideally, one would like to relax the Gaussianity assumption as well, however, it is less clear how one would proceed. Copula-based models hold some promise, but have not been very successful hitherto in high dimensions.

4. CONCLUSIONS

In my dissertation, we have proposed two novel statistical methodology. In the first chapter we proposed a Stochastic approximation based parameter estimation technique for Ensemble Kalman Filter set-up. We discussed asymptotic properties of the proposed method. Also we have done comparative simulation study with other popular approach. We have applied our method to large spatio-temporal data.

In our research, we found that our proposed approach suffers in case of very high dimensional problem. In future, we want to focus on how that problem can be solved. Also, we want to apply our proposed method more broadly to real data.

In the second chapter, we proposed a novel joint variable and covariance selection technique for the case of a nonlinear, Gaussian graphical model. In our future research, we want to investigate the theoretical properties of our approach.

REFERENCES

Anderson, J. L. (2001), "An ensemble adjustment Kalman filter for data assimilation," *Monthly Weather Review*, 129, 2884–2903.

— (2007), "An adaptive covariance inflation error correction algorithm for ensemble filters," *Tellus A*, 59, 210–224.

Anderson, J. L. and Anderson, S. L. (1999), "A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts." *Monthly Weather Review*, 127.

Aue, A. and Horváth, L. (2011), "Quasi-likelihood estimation in stationary and nonstationary autoregressive models with random coefficients," *Statistica Sinica*, 21, 973.

Aue, A., Horváth, L., and Steinebach, J. (2006), "Estimation in random coefficient autoregressive models," *Journal of Time Series Analysis*, 27, 61–76.

Bhadra, A. and Mallick, B. K. (2013), "Bayesian sparse models to analyze eQTL data," *Biometrics*.

Breiman, L. and Friedman, J. (1997), "Predicting multivariate responses in multiple linear regression (with discussion)." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 3–54.

Brown, P. J. (1980), "Aspects of Multivariate Regression (with discussion)." in *Bayesian Statistics*, eds. J. M. Bernardo, M. H. DeGroot, D. V. L. and Smith, A. F. M., Valencia University Press, pp. 247–292.

Carvalho, C. and Scott, J. (2009), "Objective Bayesian model selection in Gaussian graphical models," *Biometrika*, 497–512.

Dawid, A. (1981), "Some matrix-variate distribution theory: notational considerations and a Bayesian application," *Biometrika*, 265–274.

Dawid, A. P. and Lauritzen, S. L. (1993), "Hyper markov laws in the statistical analysis of decomposable graphical models," *The Annals of Statistics*, 1272–1317.

Dee, D. P. (1995), "On-line Estimation of Error Covariance Parameters for Atmospheric Data Assimilation ," *Monthly Weather Review*, 123, 1128–1145.

Dee, D. P. and da Silva, A. (1999), "Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: Methodology ," *Monthly Weather Review*, 127, 1822–1834.

DelSole, T. and Yang, X. (2010), "State and parameter estimation in stochastic dynamical models," *Physica D: Nonlinear Phenomena*, 239, 1781–1788.

Delyon, B., Lavielle, M., and Moulines, E. (1999), "Convergence of a Stochastic Approximation Version of the EM Algorithm ," *The Annals of Statistics*, 27, 94–128.

Dempster, A. (1972), "Covariance selection," *Biometrics*, 157–175.

Denison, D., Mallick, B., and Smith, A. (1998), "Automatic Bayesian curve fitting," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 333–350.

Evensen, G. (1994), " Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics," *J. Geophys. Res.*, 99, 10143–10162.

— (2009a), " The ensemble Kalman filter for combined state and parameter estimation ," *IEEE Control Systems Magazine*, 29(3), 83–104.

— (2009b), *Data Assimilation: The Ensemble Kalman Filter*, Springer.

Fan, J. and Gijbels, I. (1995), "Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation," *Journal of the Royal Statistical Society. Series B (Methodological)*, 371–394.

Friedman, J. (1991), "Multivariate adaptive regression splines," *The Annals of Statistics*, 1–67.

Furrer, R. (2006), "KriSp: An R Package for Covariance Tapered Kriging of Large Datasets Using Sparse Matrix Techniques," *inside. mines. edu/ rfurrer/software/KriSp*, 325, 335–336.

Furrer, R., Genton, M. G., and Nychka, D. (2006), "Covariance tapering for interpolation of large spatial datasets," *Journal of Computational and Graphical Statistics*, 15.

Gelb, A. (1974), " Applied Optimal Estimation," *The M.I.T Press*.

Gelfand, A. E. and Banerjee, S. (1998), "Computing Marginal Posterior Modes Using Stochastic Approximation," Tech. rep., University of Connecticut, Dept. of Statistics.

Gu, M. G. and Kong, F. H. (1998), "A Stochastic Approximation Algorithm With Markov Chain Monte Carlo Method for Incomplete Data Estimation Problems ," *Proceedings of the National Academy of Sciences USA*, 95, 7270–7274.

Gu, M. G. and Zhu, H. T. (2001), "Maximum Likelihood Estimation for Spatial Models by Markov Chain Monte Carlo Stochastic Approximation ," *Journal of the Royal Statistical Society, Ser. B*, 63, 339–355.

Hamill, T. M., Whitaker, J. S., and Snyder, C. (2001), "Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter," *Monthly Weather Review*, 129, 2776–2790.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall.

Holmes, C., Denison, D., and Mallick, B. (2002), "Accounting for model uncertainty in seemingly unrelated regressions," *Journal of Computational and Graphical Statistics*, 533–551.

Houtekamer, P. L. and Mitchell, H. L. (2001), "A sequential ensemble Kalman filter for atmospheric data assimilation," *Monthly Weather Review*, 129, 123–137.

Jazwinski, A. H. (1970), *Stochastic Processes and Filtering Theory*, New York and London: Academic Press.

Johns, C. J., Nychka, D., Kittel, T. G. F., and Daly, C. (2003), "Infilling sparse records of spatial fields," *Journal of the American Statistical Association*, 98, 796–806.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005), "Experiments in stochastic computation for high-dimensional graphical models," *Statistical Science*, 388–400.

Jr, P. J. R. and Diggle, P. J. (2001), "geoR: a package for geostatistical analysis," *R-NEWS*, 1, 14–18, iSSN 1609-3631.

Julier, S. J., Uhlmann, J. K., and Durrant-Whyte, H. F. (1995), "A new approach for filtering nonlinear systems," in *American Control Conference, Proceedings of the 1995*, vol. 3, pp. 1628–1632.

Kalman, R. E. (1960), " A new approach to linear filter and prediction problems," *J. Basic Eng.*, 82, 35–45.

Kalman, R. E. and Bucy, R. S. (1961), "New results of linear filtering and prediction theory ," *J. Basic Eng.*, 83, 95–108.

Kiefer, J. and Wolfowitz, J. (1952), "Stochastic Estimation of the Maximum of a Regression Function ," *Annals of Mathematical Statistics*, 3, 462–466.

Kushner, H. J. and Shwartz, A. (1984), "An invariant measure approach to the convergence of stochastic approximations with state dependent noise," *SIAM Journal on Control and Optimization*, 22, 13–27.

Lai, T. L. (2003), " Stochastic Approximation," *The Annals of Statistics*, 31, 391–406.

Lauritzen, S. (1996), *Graphical Models*, Oxford University Press, USA.

Lauritzen, S. L. (1981), " Time series analysis in 1880. A discussion of contributions made by T.N. Thiele," *International Statistical Review*, 49, 319–333.

— (2002), *Thiele: Pioneer in Statistics*, Oxford University Press.

Le Gland, F., Monbet, V., Tran, V.-D., et al. (2009), "Large sample asymptotics for the ensemble Kalman filter," .

Liang, F., Cheng, Y., Song, Q., Park, J., and Yang, P. (2013), "A Resampling-Based Stochastic Approximation Method for Analysis of Large Geostatistical Data," *Journal of the American Statistical Association*, 108, 325–339.

Liang, F., Liu, C., and Carroll, R. J. (2007), " Stochastic Approximation in Monte Carlo Computation," *Journal of the American Statistical Association*, 102, 305–320.

Liu, H. and Wang, L. (2012), "TIGER: A Tuning-Insensitive Approach for Optimally Estimating Gaussian Graphical Models," *arXiv preprint arXiv:1209.2437*.

Lorenz, E. and Emanuel, K. (1998), "Optimal sites for supplementary weather observations: Simulation with a small model," *Journal of the Atmospheric Sciences*, 55, 399–414.

Lorenz, E. N. (1996), "Predictability: A problem partly solved," in *Proc. Seminar on predictability*, vol. 1, pp. 1–18.

Mitchell, H. L. and Houtekamer, P. L. (2000), "An Adaptive Ensemble Kalman Filter," *Monthly Weather Review*, 128, 416–433.

Moyeed, R. A. and Baddeley, A. J. (1991), "Stochastic Approximation of the MLE for a Spatial Point Pattern ," *Scandinavian Journal of Statistics*, 18, 39–50.

Nicholls, D. F. and Quinn, B. G. (1982), *Random Coefficient Autoregressive Models: an Introduction*, Springer-Verlag.

Robbins, H. and Monro, S. (1951), "A Stochastic Approximation Method ," *Annals of Mathematical Statistics*, 3, 400–407.

Rothman, A., Levina, E., and Zhu, J. (2010), "Sparse multivariate regression with covariance estimation," *Journal of Computational and Graphical Statistics*, 947–962.

Smith, M. and Kohn, R. (2000), "Nonparametric seemingly unrelated regression," *Journal of Econometrics*, 257 – 281.

Sorenson, H. W. (1970), "Least-squares estimation: from Gauss to Kalman," *Spectrum, IEEE*, 7, 63–68.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer.

— (2004), "Equivalence of Gaussian measures for some nonstationary random fields," *Journal of Statistical Planning and Inference*, 123, 1–11.

Stroud, J. R. and Bengtsson, T. (2007), "Sequential State and Variance Estimation within the Ensemble Kalman Filter ," *Monthly Weather Review*, 135, 3194–3208.

Swerling, P. (1958), " A proposed stagewise differential correction procedure for satellite tracking and prediciton," *Rand Corporation*.

Welch, G. and Bishop, G. (2007), "An Introduction to the Kalman Filter," .

Whitaker, J. and Hamill, T. (2002), "Ensemble data assimilation without perturbed observations," *Monthly Weather Review*, 130, 1913–1924.

Yang, X. and Delsole, T. (2009), "Using the ensemble Kalman filter to estimate multiplicative model parameters," *Tellus A*, 61, 601–609.

Younes, L. (1988), "Estimation and annealing for Gibbsian fields," *Ann. Inst. H. Poincaré Probab. Statist*, 24, 269–294.

— (1999), "On the Convergence of Markovian Stochastic Algorithms With Rapidly Decreasing Ergodicity Rates ," *Stochastics and Stochastics Reports*, 65, 177–228.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007), "Dimension reduction and coefficient estimation in multivariate linear regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 329–346.

Zellner, A. (1962), "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," *Journal of the American Statistical Association*, 348–368.